# INTA6450 Midterm: Enron Proposal

Prakhar Yadav
pyadav67@gatech.edu

***Abstract***—In this proposal I outline how I plan to investigate and explore the Enron email corpus for the classification, detection, and prediction of wrongdoing.

## 1 REPORT

### 1.1 Definition

Wrongdoing at Enron was likely directly responsible for its collapse. Accordingly, identification of wrongdoing is a fruitful activity in mitigating shareholder losses and preserving justice in the United States. While wrongdoing may seem a nebulous concept, in the context of Enron's collapse, one may consider illegal activity as the most limited threshold of wrongdoing. One may also desire to identify wrongdoing that has not been codified into law, including misleading, deception, and bullying. These forms of wrongdoing may be driven by or be forms of collusion, which is "secret of illegal cooperation of conspiracy, especially in order to cheat or deceive others". For the purposes of this analysis, detection of precursors of illegal activity, illegal activity, as well as calls to mislead, deceive or force another to do the same, are the primary interest.

To be more specific, an example of illegal activity perpetrated by Enron is insider trading. Insider trading involves the purchase (or sale) of stocks based on proprietary (or non-public information). Insider information may have been shared amongst executives at Enron, representing a breach of fiduciary trust and compromise of market integrity.

Another example of market manipulation by Enron involves deceptive or creative accounting practices that involved the use of special purpose entities (SPEs) to hide debt. Enron over-leveraged these assets to distance itself from its investment mistakes on paper. Many losses that arose from Enron's expansion into new industries were hidden away in SPE's that aimed to show healthy financials on paper, as the financials of SPEs were either not required to be disclosed or simply not disclosed. Examples of this

An example of deception involves the use of mark to market accounting, which was especially uncommon in Enron's industry. Jeffrey Skilling introduced the practice, which involves writing "unrealized future gains from some trading contracts into current income statements, thus giving the illusion of higher current profits" (Britannica 2023). More simply, Enron represented its long term contracts and projected future earnings as current profits, demonstrating high revenues and growth that were not indicative of their current buying power or cash on hand. Perhaps even more egregiously, when a deal with Blockbuster to provide on-demand entertainment to Americans fell through, Enron continued to project profits after the deal was guaranteed a loss. When debts arose, Enron had to rely on short-term loans. Accordingly, when banks lost confidence in Enron, Enron lacked funds to pay back its debt and declared bankruptcy.

### 1.2 Signs of Wrongdoing and Strategies

With such a large dataset, some strategy will be necessary to sift insights from noise.

The first part of the strategy will likely involve some data cleansing to ensure the corpus is workable such that redacted communications and individuals do not unwontedly skew the results. For example, one may want to randomly create names to assign to some redacted emails or exclude redacted individuals emails from suspicious individuals' analysis. Another form of cleansing may involve the elimination of duplicate emails.

One may reasonably expect signs of wrongdoing as described earlier within the Enron emails corpus. Perhaps the best starting point for the identification of such signs is the search of the corpus for specific keywords. Terms that are of particular interest include, but are not limited to, "special purpose entities" and "off-the-books". Emails containing the term "quiet", "hush-hush" or "down low" may be the clearest indications of collusion or insider trading attempts. Other terms of interest may include "bankruptcy" or "debt".

Once such emails are identified and located, manual scanning or LLM summarization may enable the cultivation or finetuning of tailored sets of search terms alongside a prediction of a form of wrongdoing. One can then begin to flag these emails with confidence, as well as highlight the individuals that are most commonly in the loop during problematic or potentially problematic communications to further enrich the data. These flags can contribute to a more robust dataset and diversely featured dataset than the original unstructured dataset to produce better predictions.

Another potentially fruitful strategy involves sentiment analysis of each email. Sentiment analysis is a technique that aims to uncover emotions in text data. At minimum, one can classify documents as positive, negative or neutral, but some strategies enable the classification of more fine-grained categories for reasonable emotion detection. For instance, an email characterized by anger from a superior in response to an email characterized by disgust from a lower-ranked employee may be an indication of an attempt to bully someone to go against their morals. Accordingly and similarly, problematic emotional trends can be flagged, stored, and raised.

A final strategy that may prove fruitful is network analysis of the relationships between Enron employees. With the previous flags of email subject matter, email tone, and risky employees, one can perform network analysis to see which chains of communication are the worst perpetrators of wrongdoing. Furthermore, an aspect of the data that may be especially useful is the intensity or frequency of communication, especially as related to the release of company financial data. One would likely expect some communications from and between the executive level accounts prior to public financials releases, and would serve as a good sanity check for the viability of the strategic process. These communication spikes between executives (prior to release of quarterly earnings) may highlight low-morality networks that can be flagged for further analysis.

The culmination of the previous strategies may make it possible to automatically converge on the most problematic snippets of email threads that are indications of wrongdoing.

### 1.3 External Information

For this project, an example of external information that may be very useful is the accounts of whistleblowers or former employees. The language they use or share may be useful for the identification of search terms when parsing the Enron corpus. Any data that can be unearthed on the trading practices of Enron executives may also prove useful in the preliminary identification of insider trading. The timing of certain trades and uncharacteristic quantities or values may be preceded by a flurry of email communications. A finally useful piece of information may be public news or social media which may have clues in the form of skepticism or incredulity regarding the rise of an energy company to the top of the S&P500.

### 1.4 Methods

For Methods that I would like to leverage in pursuit of this project include Elastic Search, Pandas, SQL and in particular Databricks, which is built as extension to Spark and Hadoop. Databricks provides a unified platform to perform analysis on big data with several optimizations and conveniences baked-in. An example of this convenience is DBRX, an LLM model that can be augmented with external data for ease in building a RAG model. My hope is that during the construction of an identification or prediction framework, LLM's may make it easier to verify the contents of a flagged communication or series of communications. While some manual intervention and reading of the text may initially be required, it is possible that the DBRX LLM model can split the data into higher-confidence and lower-confidence subsets to reduce the verification load on human agents.

Elastic search, as covered in the lecture, provides a convenient and performant way to identify and locate relevant documents within the corpus that match our desired search terms. When search terms are rare, they are weighted more heavily ensuring that unique search terms return relevant results more quickly. Accordingly, when determining the best terms elastic search is likely useful.

Pandas and NLTK in conjunction with Counter, can be leveraged to cleanse data, manipulate it, and read in the text. NLTK can tokenize the email text for further manipulation and counter can count the prevalence of certain terms. It is likely certain Enron employees used some flagged terms disproportionately, and one can split the employees by identifying the individuals most responsible for the most problematic terms via measurement of the median, mean, mode and standard deviation.

For sentiment analysis, I would like to leverage the NRC word-emotion association lexicon, which provides several basic emotions such as anger, trust, and anticipation for the analysis of text. I believe that certain rations of emotion in emails, or patterns of emotions in email may emerge prior to illegal activity or collusion. By rating or ranking the prevalence of emotions over time and in email threads, may be able to train a basic decision tree classifier to predict wrongdoing. For this purpose, scikit learn provides several libraries to split the data and train a learning model. The outputs of predictions can be verified with in-sample and out-of-sample data to verify the viability of such a model. It is also possible to leverage a tree model from scratch using Quinlan's classic paper on decision tree classifiers. Decision trees are especially useful as they are easy to interpret and follow, which may make it possible to step through nodes to see the inner working of the classifier.

Finally, time-permitting, the implementation of a visual graph network may make it easier to follow wrongdoing. Hugging face offers several pretrained models that enable the generation of novel relationship graph networks which can be stored in Neo4J for querying.

To reduce errors, I plan on manually reviewing a random subset of the data, but later reviews may be assisted with LLMs to split the data into higher confidence and lower confidence analyses to reduce the need for manual intervention.

### 1.5 Problems with Methodology

An issue that may arise with any one of these strategies is false flagging, or a false positive. For example, when searching for "down low" in an email of sufficient generality, a statistical method may be unable to distinguish an email about a surprise birthday party from a more relevant or egregious email that indicates wrongdoing. Another issue with the keyword search may be plain language search may be insufficient in identifying problematic issues, as euphemisms or code may be commonplace to hide wrongdoing.

Furthermore, a basic sentiment analysis may be insufficient in providing useful flagging for data enrichment. While the aim of the proposal strategy is a cumulative ensemble effect of the several useful methodologies to enable the mitigation of noise in the dataset, there simply may be too much data to parse and process introducing some feasibility concerns given a limited timeline. An issue with more fine-grained sentiment analysis involves the neutral or professional nature of business emails as a whole. While one may hope to uncover insights, they may be limited in availability. However, by providing additional context via keywords and network analysis the sentiment data may still be useful.

Another issue may be in relation to LLMs which can hallucinate information. Reliance on an LLM may be necessary for automation, but must be exercised with caution. A potential workaround may involve the use of Retrieval Augmented Generation, a personal stretch goal, to ground LLM outputs. A final issue in relation to machine learning methods may involve overfitting and the time series nature of the data. If training on 'future' data or emails, the predictions of a model may be skewed, biased, and unreliable. The model may essentially be peeking into what happened rather than producing a clean prediction.

## 2 REFERENCES


1. Bondarenko, P. (2024, July 29). Enron scandal. Encyclopedia Britannica. https://www.britannica.com/event/Enron-scandal
2. Collusion, N. (2023). *Oxford English Dictionary*. Oxford University Press. https://doi.org/10.1093/OED/9192217621
3. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465.
4. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251