Enron Email Machine Learning Project – Project Planning Notes

Executive Summary:

This project leverages the Enron Email Dataset to explore how machine learning and natural language processing can identify communication patterns that signal potential organizational risk or misconduct. Using a structured data science process (CRISP-DM), the analysis focuses on cleaning and transforming raw email data into features such as sentiment, topic categories, and message frequency. Multiple machine learning and deep learning models will be evaluated to classify and visualize email behaviors that may indicate anomalies or emerging risks. The results will provide insights into how automated text analytics could support compliance, fraud detection, and early warning systems in corporate environments. Findings and recommendations will be communicated through interactive visualizations and a presentation tailored for management decision-making. The analysis will process approximately 500 000 emails to identify outlier communication clusters by sentiment and topic shift.

Problem Statement / Research Objectives:

The collapse of Enron highlighted the critical importance of internal transparency and early detection of unethical communication behavior. This project seeks to explore whether machine learning and natural language processing can be used to identify patterns in internal emails that may indicate fraud, collusion, or emerging risk within an organization. By analyzing the Enron Email Dataset, the project will examine how sentiment, message frequency, and topic distribution correlate with abnormal communication activity.

The primary research objectives are to:

1. Engineer a set of text-based and metadata features from Enron's internal emails that capture communication behavior.

2. Develop and compare multiple machine learning models to detect potentially anomalous or high-risk communication.

3. Visualize the findings and assess how such an approach could support compliance and corporate risk monitoring in real-world organizations.

Exploratory Data Analysis (EDA):

The Enron Email Dataset contains a large collection of internal communications between Enron employees from 1998 to 2002. The data includes sender and recipient information, timestamps, subject lines, and email body text. For this analysis, a subset of the dataset will be sampled and preprocessed to remove duplicates, system messages, and corrupted files.

The initial exploration will focus on understanding the structure, completeness, and key patterns within the data.

Descriptive statistics and visualizations will be used to analyze:

- The total number of emails, senders, and recipients
- Distribution of emails over time (to identify periods of abnormal communication)
- Top senders and recipients by volume
- Average message length and frequency by department or individual
- Word frequency and key topics using term-frequency and n-gram analysis

- Overall sentiment distribution across emails and how it changes over time

Visualization tools such as Python's Matplotlib, Seaborn, and WordCloud libraries will be used to highlight trends, while network graphs will help illustrate the flow of communication between key individuals. The findings from this stage will help guide feature engineering and model selection by identifying variables that correlate with unusual or high-risk communication patterns.

Data Preparation and Feature Engineering:

The Enron Email Dataset contains thousands of raw text files with inconsistent formatting, missing values, and embedded system information. Before developing predictive models, the data will be cleaned, standardized, and transformed into structured features suitable for machine learning.

The preparation process includes:

1. Data Cleaning: Remove duplicate and empty emails, strip out headers, footers, and non-text artifacts such as "From," "To," and "X-FileName" fields. Convert all text to lowercase and remove punctuation, stop words, and extra whitespace.

2. Tokenization and Lemmatization: Split sentences into words (tokens) and normalize them to their root forms using NLTK or spaCy.

3. Feature Engineering:

    - Text Features: Generate TF-IDF vectors and sentiment scores for each email.
    - Metadata Features: Include sender and recipient domains, email length, time of day sent, and message frequency per user.
    - Topic Features: Use Latent Dirichlet Allocation (LDA) or transformer-based embeddings to extract thematic patterns.

4. Variable Transformations: Apply standard scaling or normalization where applicable, particularly for numeric variables such as message counts or lengths.

5. Feature Selection: Evaluate which engineered variables provide the strongest signal for detecting anomalies or risk-related communication.

The resulting dataset will combine text embeddings with numerical features into a structured dataframe ready for model development. All preprocessing steps will be modularized in Python scripts for reproducibility and later deployment.

Findings and Conclusions:

The machine learning analysis compared three classification algorithms:

1. Logistic Regression
2. Random Forest
3. Gradient Boosting

Using TF-IDF text features and email-level metadata to determine whether messages were internal (sent from an Enron address) or external. All three models performed well, achieving F1-scores above 0.93.

The Random Forest model achieved the best result with an F1-score of 0.94, indicating a strong balance between precision and recall. The Gradient Boosting model produced nearly identical results, suggesting that ensemble-based approaches capture the non-linear relationships inherent in textual and metadata features more effectively than linear methods.

The results confirm that supervised learning can accurately classify communication patterns within the Enron dataset. These insights could be extended to support fraud detection, compliance monitoring, or insider risk management applications by analyzing tone, frequency, and sender-recipient relationships at scale.

Lessons Learned:

Working with the Enron dataset revealed the importance of thorough preprocessing and feature engineering in text-based machine learning projects. Parsing and cleaning over a million raw emails required handling missing metadata, normalizing inconsistent text formats, and constructing meaningful features such as sentiment, body length, and recipient count. Once the text was vectorized with TF-IDF, it became clear that even simple lexical patterns could accurately distinguish internal and external communications.

Model experimentation showed that ensemble methods (Random Forest, Gradient Boosting) consistently outperformed linear classifiers. This suggests that relationships within email language—such as tone, phrasing, and domain-specific vocabulary—are non-linear in nature. Additionally, reducing dimensionality and using cross-validation helped control overfitting while maintaining interpretability.

The project also highlighted the scalability challenges of large, unstructured datasets. Future iterations would benefit from distributed processing or deep learning architectures (e.g., LSTM or transformer-based models like BERT) to capture richer semantic meaning. Overall, this project provided practical experience in text mining, end-to-end pipeline design, and model comparison using real-world data.