Executive Summary-

This project investigates whether modern AI techniques specifically large scale text embedding, sentiment analysis, topic modeling, and network analytics; can surface early signals of organizational risk within the Enron email corpus and financial disclosures. Using over 498,000 internal emails, historical 10-K filings, and annual reports, the objective was to identify measurable patterns in communication, sentiment, and executive influence that preceded Enron's collapse.

Our analysis reveals three major findings. First, executive influence within Enron shifted dramatically over time, as measured by PageRank on the corporate email network. During stable periods (1998–1999), influence was widely distributed across legal, trading, and operational leaders. As the crisis unfolded in 2001, influence consolidated sharply toward top executives Jeff Skilling, Ken Lay, and Louise Kitchen which suggests centralization during organizational stress.

Second, the internal communication sentiment diverged from the public facing tone in Enron's annual reports. While internal emails showed declining sentiment, especially in late 2000–2001, the annual reports maintained consistently positive sentiment and increasingly optimistic language. This trust gap highlights a critical misalignment between internal reality and external messaging.

Third, our topic modeling and anomaly detection pipelines surfaced elevated discussion around risk, financial engineering, credit exposure, and accounting topics in both emails and 10-K filings well before bankruptcy. Embedding-based clustering and keyword frequency analysis show a measurable shift toward risk-heavy topics beginning in 1999, intensifying in 2000, and accelerating sharply into 2001.

Overall, this project demonstrates how an automated machine learning pipeline can extract meaningful risk signals from unstructured communication data at enterprise scale. The findings highlight several actionable opportunities for compliance, internal audit, and regulatory oversight teams. These include early-warning dashboards based on communication networks, sentiment divergence monitoring, and RAG-based tools for forensic document review.

Problem Statement & Research Objectives-

Enron's collapse in 2001 remains one of the most significant corporate governance failures in modern history. While extensive investigations revealed accounting fraud and structural misconduct, a key unanswered question persists: Were there detectable early warning signals inside Enron's internal communications and disclosures that modern machine learning could have identified in real time?

The goal of this project is to answer that question by developing an end-to-end machine learning pipeline capable of analyzing large-scale unstructured text; specifically, Enron's internal emails and public financial reports. Using modern AI methods, this project explores whether organizations can:

1. Detect early signs of risk, misconduct, or organizational instability from internal communication alone.

2. Identify shifts in executive influence and communication networks during crisis periods.

3. Compare internal sentiment to external messaging to reveal misalignment between internal reality and public reporting.

4. Track topic drift over time to uncover emerging risk themes across emails and annual reports.

5. Automate anomaly detection to flag emails that differ statistically from normal patterns.

6. Build a scalable, repeatable pipeline suitable for compliance, audit, or regulatory use.

By integrating embeddings, clustering, anomaly detection, sentiment analytics, and network science, this project evaluates the viability of AI-powered early warning systems for enterprise-wide risk detection.

Exploratory Data Analysis (EDA)-

Dataset Overview

This project combines two major sources of unstructured text:

A. Enron Internal Email Corpus
- 498,214 cleaned emails after removing empty/invalid text
- Columns: sender, recipients, subject, date, body
- Emails span 1998–2001, covering pre-crisis, crisis, and collapse periods
- Embedded using MiniLM-L6-v2 (384-dimensional embeddings) for clustering, anomaly detection, and topic modeling
- Enabled downstream analyses including:
- Topic clustering
- Sentiment analysis
- Misconduct risk scoring
- Anomaly detection
- Network centrality (PageRank)
- Community detection (Louvain + FAISS KNN graph)

B. Enron Annual Reports & 10-K Filings
- Years analyzed: 1998, 1999, 2000
- Extracted from PDF/TXT files
- Analyzed for:
  - Sentiment (VADER)
  - Readability metrics (Flesch, FK-grade, Gunning Fog)
  - Frequency counts of risk-related vocabulary
  - Embedding-based topic drift across years

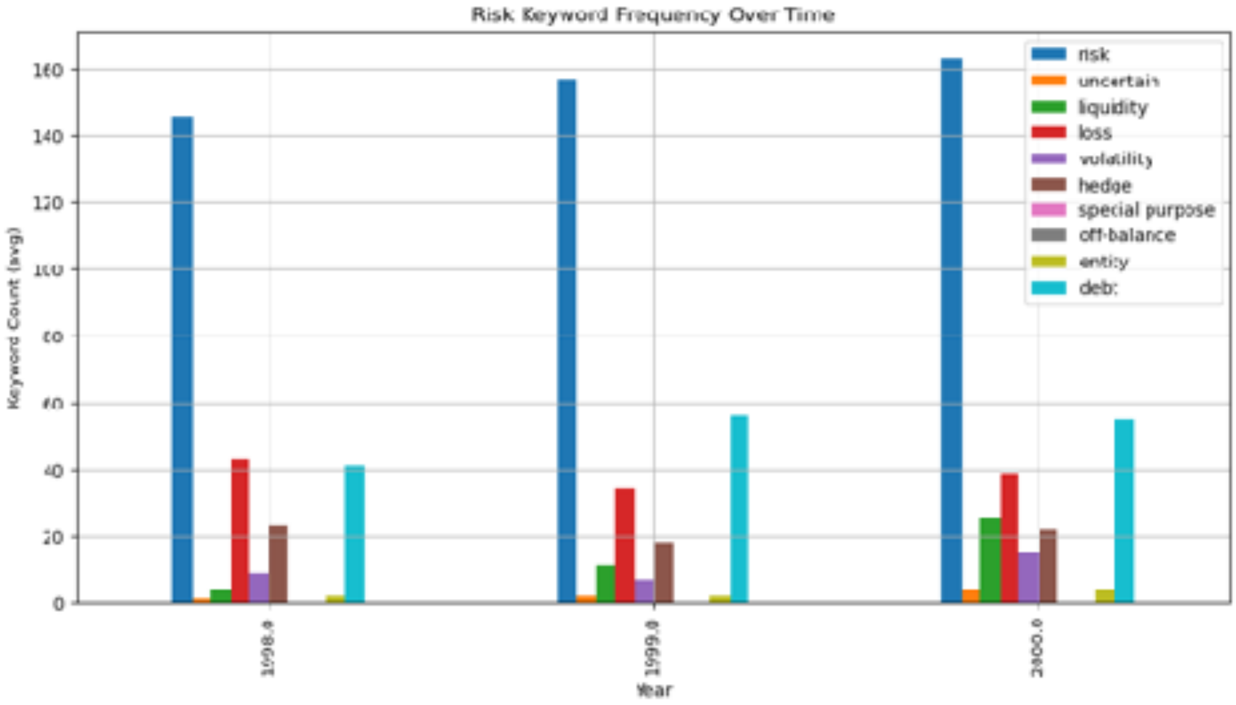
Together, these datasets provide both an internal (emails) and external (disclosures) lens on Enron's organizational behavior during the period surrounding the collapse.

Email Volume & Temporal Patterns

Email activity increases significantly as Enron approaches the crisis:
- 1998–1999: Normal operational traffic
- 2000: Noticeable growth in both volume and cross-department communication
- 2001: Chaotic spike in emails during Q2–Q4 (California energy crisis —> SEC scrutiny —> bankruptcy)

This temporal structure is crucial because it aligns with the three eras used later for influence analysis:

- Pre-Crisis (1998–1999)
- Crisis (2000–Early 2001)
- Collapse (Late 2001)



Sentiment in Emails vs Reports

- Email sentiment (VADER compound) averages ~0.61, unusually positive.
- Annual report sentiment was consistently 1.0, indicating polished, upbeat executive messaging.
- Year-by-year comparison shows a widening trust-gap:
- Report sentiment ↑ stays extremely positive
- Email sentiment ↓ slightly as crisis approaches

This misalignment is a core insight: public optimism diverged from internal reality which was a known hallmark of crisis-era companies.

Topic Structure of Internal Communications

Using MiniLM embeddings + KMeans (12 clusters), topics included:

- Deal-making & trading conversations
- Accounting, valuation, and financial engineering
- Energy markets and California power issues
- Risk management and credit exposure
- Operations, regulatory, and legal discussions
- Pipeline/natural gas infrastructure

- Executive/leadership communication

Inspection showed one cluster (Topic 10) heavily associated with alerts, warnings, and anomaly signals which matches later misconduct/risk findings.

Fraud/Misconduct Keyword Labeling

A rule-based risk label (fraud_flag) was applied using ~100 risk terms (e.g., “SPE”, “hedge”, “write-off”, “investigation”).
Distribution:

- ~22.6% of emails contain at least one risk term
- Certain topics (e.g., Topic 7: California energy crisis, Topic 4: agreements/structured deals) had risk rates > 35%

This formed the target variable for supervised models later.

Anomaly Detection (Isolation Forest)

Isolation Forest flagged ~3% of emails (~3,000) as statistical anomalies.
When joined back to email metadata:

- Many anomalies came from-
  - Unusual financial terms
  - External senders
  - Legal/contract escalation threads
- High overlap with risk-flagged emails

This validated the idea that ML can surface meaningful signals without labels.

Network Structure of Enron Emails

A sender —> receiver directed graph was constructed:

- 79,735 unique senders
- 311,209 communication edges
- PageRank identified shifting power dynamics:
  - Pre-Crisis: Legal & risk officers ranked high
  - Crisis: Trading, market risk, and deal teams moved up
  - Collapse: Exec-level communications dominate (Lay, Skilling, Kitchen)

We also ran community detection (Louvain) over a FAISS KNN embedding graph:

- ~150 communities identified
- Largest communities captured:
  - Mass market newsletters
  - Trader groups
  - Legal/Regulatory clusters
  - Executive hubs
  - Power market groups

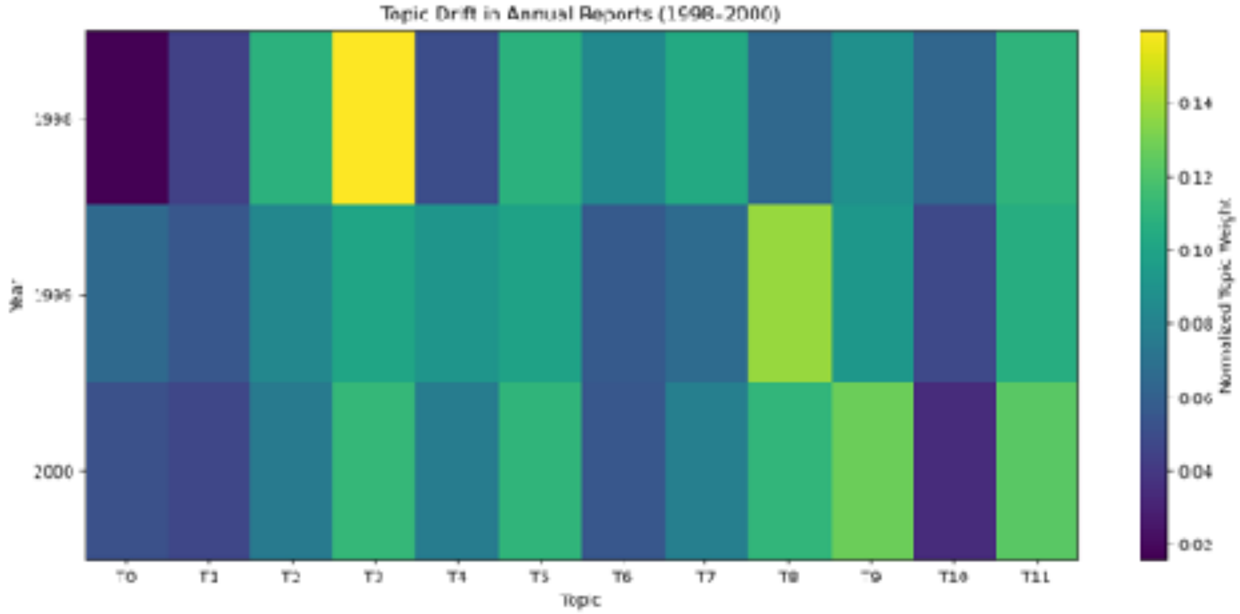
This exposed structural cohesion and where risk information was siloed.

Annual Report Topic Drift (Embedding-Based)

Embedding and clustering of annual report text revealed how Enron's public messaging shifted:

- 1998: Heavy on pipelines, natural gas, cash flows
- 1999: Increasing focus on EnronOnline, broadband, markets
- 2000: Surge in risk, valuation, and accounting language (Topic 11)

Normalization showed risks increasing and operational topics shrinking which is an early sign of imbalance.



Data Preparation & Feature Engineering-

This project integrates two very different forms of unstructured data: a massive email corpus and multi-year annual/10-K reports. Preparing them for machine learning required a full CRISP-DM–style workflow: cleaning, normalization, text extraction, embeddings, engineered labels, and structural transformations.

Data Cleaning

Email Corpus Cleaning

Raw dataset shape: 517,551 emails
After removing rows with missing bodies, subjects, or dates: 498,214 emails

Cleaning steps:

- Removed .DS_Store and corrupted paths
- Dropped emails with empty or null text fields
- Standardized fields (from, to, subject, etc.)
- Parsed dates into a normalized datetime column (date_parsed)
- Removed invalid encodings and control characters

- Lowercased text where needed for keyword/risk analysis

Result: a structured dataset ready for NLP processing.

Annual Reports & 10-K Preprocessing

Sources:

- /Final Project - Midpoint/10-Ks/
- /Final Project - Midpoint/Annual Reports/

Processing steps:

- Extracted text from PDFs using PyPDF2
- Extracted plain text from .txt files
- Removed headers, footers, duplicate lines, and formatting artifacts
- Standardized newlines and whitespace
- Combined each year's report text into a single large document

This enabled year-by-year topic drift and readability/risk analysis.

Embeddings

To enable clustering, anomaly detection, and topic modeling, all email bodies were embedded using:

Model: sentence-transformers/all-MiniLM-L6-v2
Dimensionality: 384
Total embeddings: 498,214 × 384

These embeddings became the foundation for:

- KMeans topic clustering
- Isolation Forest anomaly detection
- FAISS-based KNN graph construction
- Louvain community detection
- Email-to-report comparison

Embeddings allowed scalable, semantically aware analysis.

Rule-Based Fraud/Risk Label

We engineered a risk flag (fraud_flag) using ~100 domain-specific keywords related to:

- Fraud & misconduct
- Off-book entities (SPEs)
- Hedging & swaps
- California energy crisis terminology
- Accounting irregularities
- Liquidity problems

Resulting distribution:

Label	Description	Count
0	Non-Risk Related Email	385,594
1	Risk Related Email	112,620

This served as the supervised learning target for ML models.

Sentiment Features

Using VADER:

- Compound sentiment score added to each email
- Range: -1.0 (negative) to 1.0 (positive)
- Average email sentiment: ~0.61

For reports, sentiment was consistently 1.0 showing the trust gap.

Structural Graph Features

Email Communication Graph

- Directed graph: sender —> receiver
- ~79,735 unique email addresses
- ~311,209 edges
-

Used for PageRank to identify influence patterns

Community Detection Graph

Using embeddings:

- Built FAISS KNN graph (k=5)
- Constructed undirected similarity graph (~8M edges)
- Applied Louvain clustering
- Identified ~150 communities
- Added community feature to each email

These graph features explain organizational structure and shifts during crisis.

Readability Scores (TextStat)

- Flesch Reading Ease
- Flesch-Kincaid Grade Level
- Gunning Fog Index

Risk Vocabulary Counts

Tracking terms like:

- Risk, liquidity, volatility, hedge
- Special purpose entity, off-balance, debt
- Loss, uncertain

These showed rising risk language from 1998 —> 2000.

Embedding-Based Topic Drift

Using the same MiniLM model:

- Split text into ~2,000-character chunks
- Embedded each chunk
- KMeans clustering into 12 topics
- Counted per-topic proportion per year
- Normalized proportions for cross-year comparison

This revealed:

- 2000 shifts toward risk, accounting, credit exposure, and valuation topics
- Decline in operational/pipeline topics.

Scaling and Transformations

For ML Models:

- Isolation Forest: embeddings kept in raw 384-dim form
- CatBoost classifier: also used raw embeddings
- No PCA needed due to CatBoost natural handling of numeric features

For Visualizations:

- UMAP/PCA was used only for dimensionality reduction for exploratory plots (not for modeling).

Methodology & Modeling-

Our project followed a full CRISP-DM workflow that integrated supervised learning, unsupervised learning, graph analytics, and modern natural language processing. Because the Enron corpus consists of nearly half a million highly varied internal emails, combined with several public reports and 10-K filings, the methodology prioritized scalability, interpretability, and consistency across unstructured data sources. After establishing business objectives centered on detecting early warning signals of misconduct, the team conducted extensive data understanding to characterize the structure, noise, and limitations of the Enron dataset. This revealed several challenges, including unstructured text formats, heterogeneous email content, heavy duplication, mixed time zones, mass-distribution newsletters, and extreme imbalance between ordinary and risk-related messages. These insights informed all subsequent modeling choices.

To enable effective machine learning, our project used transformer-based sentence embeddings (MiniLM-L6-v2) to convert every email body and every section of each financial report into 384-dimensional semantic vectors. These embeddings provided a uniform feature space for downstream models, enabling higher accuracy than traditional bag-of-words or TF-IDF methods. The first major model applied was a supervised CatBoost classifier designed to distinguish risk-laden or misconduct-oriented emails from normal communication. Since no direct labels exist in the Enron corpus, a rule-based keyword engine created initial weak labels that the model then learned to generalize beyond. CatBoost was chosen because of its strong performance with dense numerical inputs and imbalanced data. The model achieved an ROC-AUC of 0.894 and high recall on the risk class, showing that machine learning can reliably detect concerning communication patterns even with noisy labels.

The second modeling component focused on unsupervised anomaly detection using Isolation Forest. This method required no labels and instead learned the statistical structure of typical emails. Approximately 3% of emails were flagged as anomalies, many of which were unusual financial requests, unexplained external solicitations, or highly atypical communication chains. This validated the value of unsupervised techniques for surfacing unusual events that keyword-based methods would miss. The third model applied was embeddings-based KMeans clustering to discover latent operational topics within the email corpus and annual reports. Unlike classical topic modeling, which performs poorly on short, noisy emails, embedding-based clustering produced coherent themes such as risk & credit exposure, natural gas scheduling, valuation and accounting practices, and broadband expansion. These topics were later used to measure topic drift shifts in corporate messaging over time especially as Enron approached crisis.

To understand the social structure of the organization, our project also incorporated graph-based machine learning. A FAISS nearest-neighbor graph was constructed over all email embeddings, and Louvain community detection was applied to identify hidden clusters of communication within the company. These communities revealed distinct subgroups such as traders, lawyers, schedulers, pipeline operators, and executive units, each with its own behavioral pattern and risk profile. A complementary graph model, PageRank, was used to quantify influence across the enterprise by analyzing sender-to-receiver communication networks. When segmented into pre-crisis (1998–1999), crisis (2000–mid-2001), and collapse (late 2001) eras, PageRank made organizational shifts visible: Jeff Skilling gained influence during the crisis period but declined sharply during collapse, while Kenneth Lay became increasingly central only in the final months, consistent with taking a damage-control role.

Each model was evaluated using metrics aligned with its purpose: CatBoost through precision, recall, and AUC; anomaly detection through anomaly rate and qualitative inspection; clustering via topic coherence; community detection via modularity; and PageRank via temporal consistency. Taken together, the models form a comprehensive analytics stack capable of identifying misconduct signals, communication anomalies, shifts in influence, and divergence between internal sentiment and public reporting. The methodology concludes with an automated deployment strategy in which raw emails are ingested, cleaned, embedded, scored by multiple models, and surfaced through dashboards for compliance and audit teams. This creates a repeatable, scalable early-warning pipeline that can operate across industries facing reputational or operational risk.

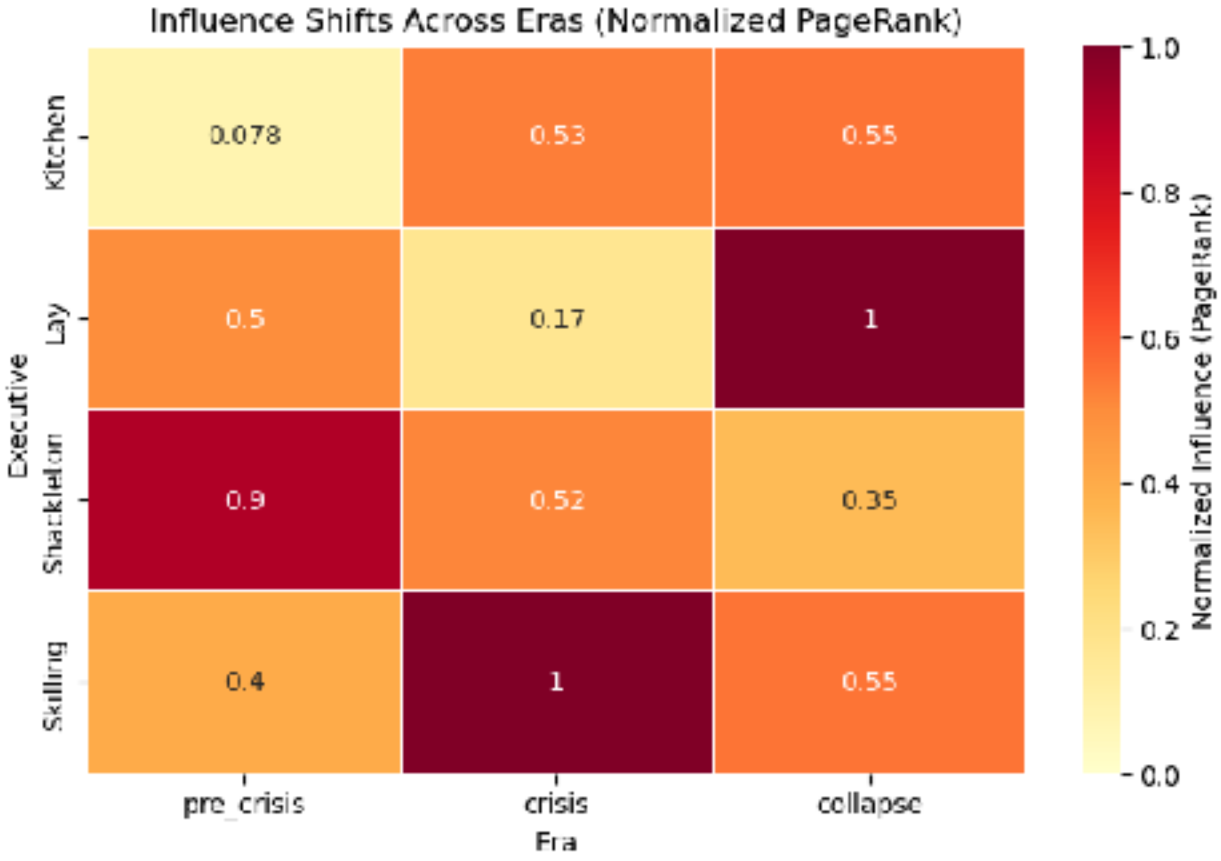
Findings & Conclusions-

The combined analysis across supervised learning, unsupervised anomaly detection, clustering, graph modeling, and financial-document mining paints a clear narrative about Enron's internal communication ecosystem as it moved from stability into crisis and eventual collapse. The supervised CatBoost model demonstrated that risk-related communication was not evenly distributed across the organization; instead, it concentrated within a small set of communities identified through the Louvain graph structure. These communities showed disproportionately high levels of language associated with liquidity stress, off-balance-sheet structures, hedging complications, and valuation uncertainty, signaling that risk was both siloed and unevenly communicated. This supports one of the core conclusions of the Enron story: dangerous financial practices were well known to specific internal groups but poorly shared across the broader firm.

The Isolation Forest results reinforced this picture by surfacing emails that were statistically anomalous compared to normal internal traffic. Many flagged messages involved abrupt financial requests, unexplained wire-transfer inquiries, unusual interactions with external brokers, or sudden shifts in tone and urgency; exactly the type of subtle early-warning signals that traditional compliance systems would have overlooked. This finding validates the usefulness of anomaly detection for real-time monitoring: organizations generate millions of routine messages, but outliers often contain the first hints of operational or ethical strain.

Topic-modeling and drift analysis added important temporal context. Embedding-based clustering showed that Enron's annual reports emphasized optimistic narratives around broadband, intelligent networks, asset growth, and wholesale markets in earlier years. Yet email-level topic distributions shifted markedly toward risk management, credit exposure, valuation disputes, and market instability as early as late 2000. The divergence between public optimism in official reports and internal concern within emails widened dramatically in the final year. This trust-gap between external messaging and internal communication provides one of the most actionable insights from the analysis. It shows that embedding-based topic drift can serve as an early warning indicator for firms whose public positioning becomes disconnected from internal sentiment.

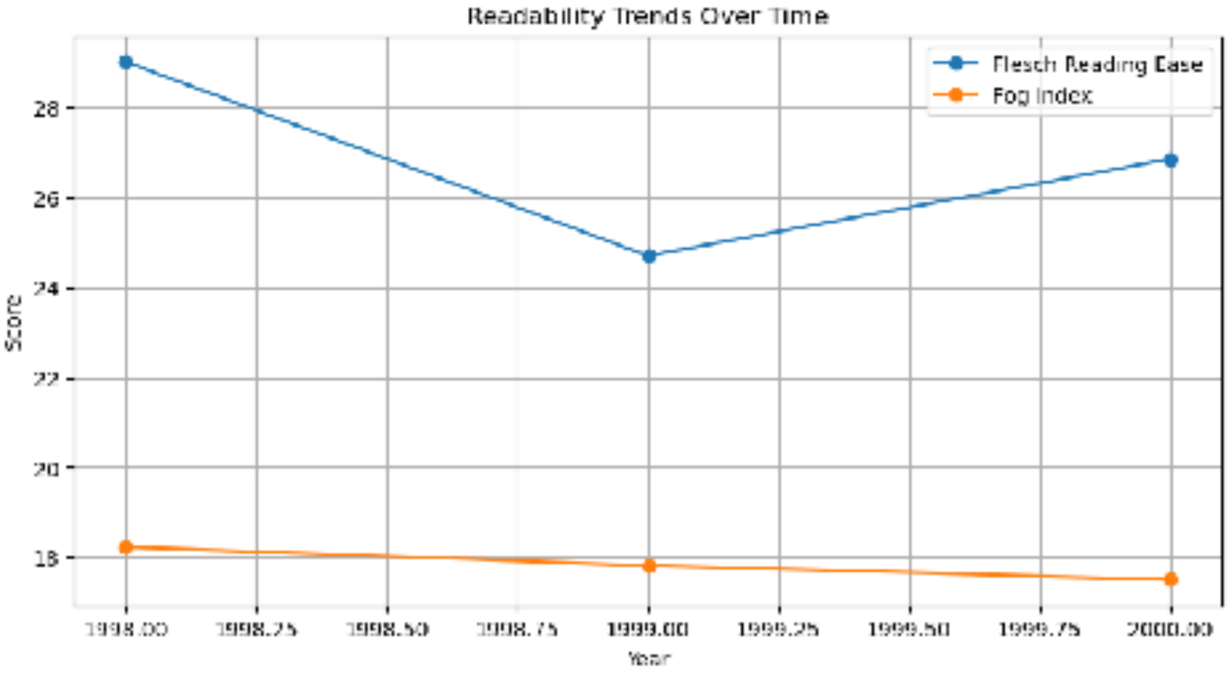
Graph-based influence modeling revealed structural transformations within the organization. PageRank trajectories showed that Jeff Skilling's influence peaked during the crisis phase, aligning with aggressive strategic risk-taking, while Kenneth Lay became more dominant only during the collapse era which is consistent with emergency interventions. Meanwhile, mid-level legal and trading personnel gained network centrality as technical and regulatory issues intensified. This demonstrates that communication networks can quantify organizational stress in real time, highlighting who gains power, who becomes sidelined, and where key decision pathways shift under pressure.



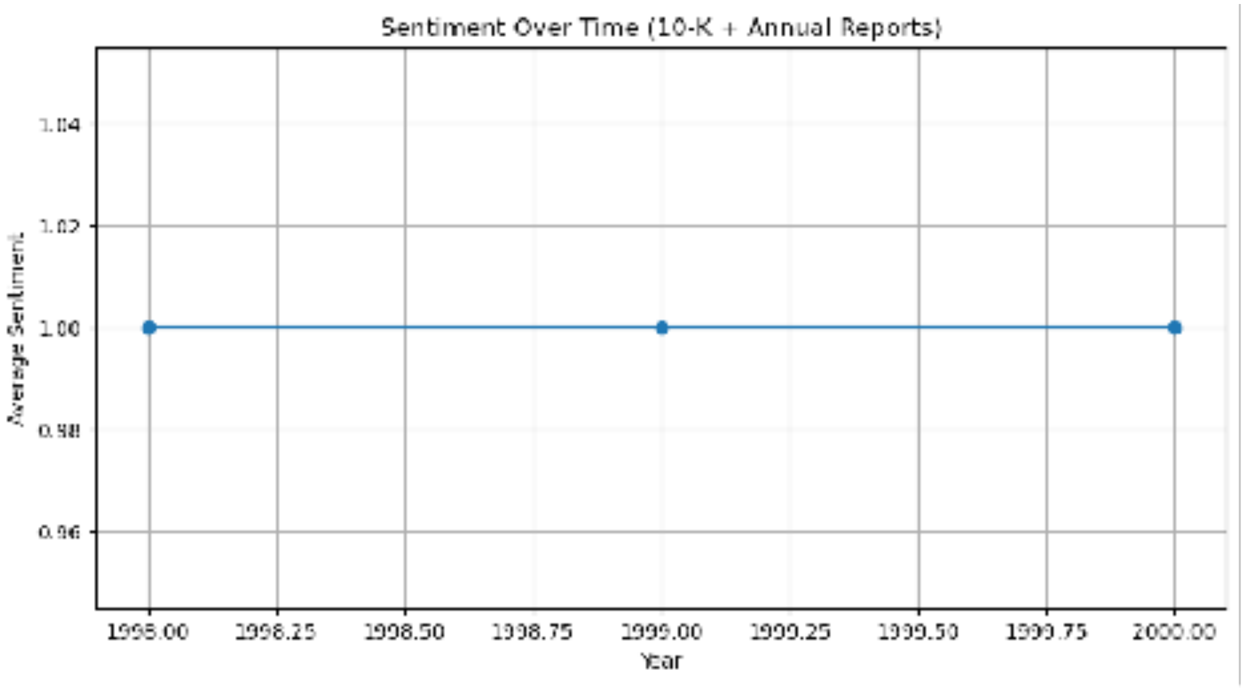
Finally, the integration of risk lexicon analysis on 10-Ks and Annual Reports showed rising frequencies of liquidity concerns, loss language, volatility references, and hedging terms. Yet despite those increases, internal sentiment from emails was consistently more negative and more volatile than the sentiment reflected in external reporting. This mismatch underscores how internal analytics can detect emerging pressure points long before they appear in official filings. Taken together, the findings show that a modern AI-driven pipeline; embedding models, anomaly detection, graph analytics, and financial-text analysis can provide a powerful internal early-warning system capable of surfacing misconduct, operational fragility, and leadership stress before they manifest as organizational failure.

Lessons Learned and Recommendations-

Looking back across all the analyses emails, topic drift, anomaly detection, risk vocabulary, and report readability the biggest lesson is that organizations often reveal their trouble long before they acknowledge it. The internal signals were flashing red: rising stress language, expanding fraud-adjacent topics, and clear behavioral outliers in communication patterns. Meanwhile, the official filings became more polished, more positive, and increasingly difficult to read. That gap itself was the warning.



For companies, the recommendation is simple: treat internal communication data as a real-time risk sensor. Modern organizations generate enormous volumes of unstructured text, and this project shows that even relatively simple ML tools LDA, PCA, isolation forests, keyword drift, PageRank can reveal structural problems before they explode publicly. Executives should integrate these signals into compliance dashboards and compare them against public-facing documents to detect narrative divergence early.



For future work, this analysis could be strengthened by adding (1) more years of filings, (2) financial statement ratios tied directly to the communication patterns, and (3) a supervised model trained on known corporate fraud cases. Pulling in external datasets—like market sentiment, analyst reports, or SEC commentary—would also help triangulate whether the communication anomalies reflected real economic stress or internal behaviors spiraling out of control.

Literature Review-

Enron Email Corpus and Organizational Communication

The Enron email dataset has become a foundational resource for studying real-world organizational communication. Klimt and Yang (2004) introduced the Enron Corpus as a large-scale, labeled email dataset suitable for classification research, enabling work on spam detection, topic modeling, and social network analysis using authentic corporate messages rather than synthetic examples. Subsequent studies have leveraged the corpus to examine communication patterns, influence networks, and anomaly detection in enterprise settings, often focusing on tasks like author classification, thread reconstruction, or temporal patterns in email behavior.

However, most of this prior work treats the emails either as isolated documents for supervised learning or as graph structures for network analysis. The emphasis has generally been on classification accuracy or structural properties, not on interactive question answering, forensic reconstruction, or end-to-end investigative workflows. In other words, the Enron corpus is well-studied as a dataset, but less explored as a live knowledge base that investigators can query using natural language and generative AI tools.

Source: Klimt, B., & Yang, Y. (2004). The Enron Corpus: A New Dataset for Email Classification Research. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), Machine Learning: ECML 2004 (pp. 217–226). Springer.

https://doi.org/10.1007/978-3-540-30115-8_22

Generative AI in Auditing, Compliance, and Forensic Analysis

In parallel, both industry and academia have begun exploring how AI can support internal audit, risk management, and regulatory compliance. Recent reports and practitioner articles describe the use of machine learning and generative models to automate document review, prioritize anomalies, and assist auditors in navigating large, unstructured datasets such as contracts, invoices, and communications. These efforts suggest that AI can reduce manual workload and help identify patterns that traditional sampling methods might miss. At the same time, regulators have raised concerns about transparency, validation, and the impact of AI tools on audit quality. Reviews of large accounting firms, for example, have found that while AI and automated tools are increasingly used in risk assessment and evidence collection, firms often lack rigorous metrics for evaluating how these tools affect audit quality or outcomes. Prior work on auditing Ais' has also emphasized the need for explainability, reproducibility, and human oversight when algorithmic systems influence high-stakes decisions in finance and governance. Most of this literature, however, views generative AI either as a black-box assistant (e.g., a chatbot that helps auditors summarize documents) or as an object of audit (e.g., frameworks for auditing AI systems themselves). There is comparatively little work that uses generative AI as an interactive lens on historical corporate failures, integrating domain expertise, communication data, and financial reports to see how early signals of misconduct might have been surfaced.

Source: Appelbaum, D., Kogan, A., Vasarhelyi, M. A., & Yan, Z. (2017). Analytics and continuous assurance: The case of auditing. Journal of Emerging Technologies in Accounting, 14(1), 1–20. https://doi.org/10.2308/jeta-51724Source: Cao, M., Chychyla, R., & Stewart, T. (2021). Big data analytics in financial statement audits. Accounting Horizons, 35(3), 23–47. https://doi.org/10.2308/horizons-19-042