Cyber Security and Artificial Intelligence
Email Gateway Data Protection

Adi Dereviani Prager, 305674731
Ron Kozitsa, 312544240

1 Introduction

Email gateway data protection is crucial for safeguarding sensitive information within an organization. By deploying an email gateway data protection system, businesses can ensure that their internal communications are secure and comply with company policies. This report will explore the implementation of an email gateway that focuses specifically on monitoring and controlling outgoing emails to prevent the inadvertent or unauthorized sharing of sensitive data.

1.1 Motivation

In an organization with numerous employees, each with varying levels of access to confidential information, the risk of data leaks or breaches through email is significant. For instance, an unintended email containing the CEO's salary details could be sent organization-wide by a new HR employee, or confidential financial reports could mistakenly reach individuals who lack the requisite clearance. These scenarios underscore the need for a robust system to manage and secure email communications.

Our email gateway data protection serves this purpose by enforcing policies that analyze outgoing emails to detect and block content that violates established rules, while permitting the transmission of compliant messages. This not only helps in maintaining the confidentiality of sensitive information but also in enforcing organizational policies effectively.

1.2 Implementing an email gateway data protection comes with many challenges

Building an email gateway data protection requires us to handle many challenges:

- Block Emails that are violating the policy, but also make sure to not block compliant ones (low to zero false positive rate).
- The implementation of the emails analysis must be quick, as users cannot wait long for emails to be sent, this requires our analysis model to be "light" and limits us from using efficient ML and DL models that have a long run time and are considered "heavy".
- The gateway needs to grasp the relevant information from emails of potentially hundreds of thousands of users, each with different email writing styles, vocabulary, and writing abilities. The gateway needs to be able to determine what the sender is talking about and make a correct decision.

1.3 Commonly used methods for implementing email gateway data protection

The common methods that are used for implementing email gateway data protection include a number of NLP techniques, such as rule-based (Regex) and dictionary-based word matching, as well as various

ML and DL techniques (Lemmatization, Stemming, NER, CNN and more).

Our research has been proven useful for guiding us with a solution.

By exploring the various methods and their pros and cons, we have learned that in some cases, there are methods that can be more accurate than others (ML approaches such as CNN vs dictionary-based) in getting our goal, but it comes with a heavy cost of computational time and resources. Therefore, we have decided to divide our approach, and create a mixture of both, not to create a "light and stupid" model and not a "smart but computationally over-expensive" model.

1.4 Our approach for the solution

We have decided to use a simple Rule-based (Regex) approach to find sensitive data that has a static known format (Emails, SSN etc.), and use an ML approach to find more dynamic data that can come in many forms, such as topics and sentiment analysis.

2 Related Work

In the field of email gateway data protection, a combination of techniques is typically employed to ensure comprehensive protection against various threats such as phishing, malware, spam, and data leakage. Though, for our usage - phishing, malware, and spam are irrelevant so we can filter out methods used to protect against those threats.

These are some of the methods used today for email gateway data protection:

- **Rule-Based approach:** Rule-based approach such as regular expressions (Regex), to identify and block or flag emails that match certain patterns indicative of malicious content or policy violations. Rules can be crafted based on known signatures of specific compliance requirements.
- **Machine Learning (ML) and Neural Networks:** Machine learning techniques, including neural networks, are increasingly utilized in email security to augment rule-based approaches. ML algorithms can analyze large volumes of data to detect patterns and anomalies that may not be captured by static rules alone.
  - **Convolutional Neural Networks (CNN):** CNNs can be employed for tasks such as detecting the content of emails to identify indicators of malicious intent.
- **Content Filtering:** Content filtering techniques are used to inspect the content of emails for sensitive information or policy violations. This can include scanning for keywords related to confidential data, such as credit card numbers or personal identification information, and enforcing encryption or blocking policies accordingly.

2.1 Do these methods fit the problem we are facing?

- **Rule-based approach - (REGEX):** We found that it is significantly more effective to use Rule-based Filtering (Regex) to capture sensitive information and patterns that are easily found using known formats. For example, email addresses. Emails have a worldwide accepted format, so it is only trivial to construct a regex to look for that format, rather than trying to use an ML approach to find if an email address is present in the context of the email.
- **Machine Learning (ML) and Neural Networks (CNN):** For our problem, we found that using LDA would give us the best results for getting the topic of the email message. Using Rule-based was just not good enough, as there is no finite set of rules that covers what is considered

'legal' or 'finance', and even if there is, we have no way to find it, so an ML approach is better for us here.

- **Content Filtering:** This method is not useful for us, as we are required to block or allow an email from being sent, and not tamper with its content.

3 Dataset Description and EDA

3.1 The Dataset

The dataset used in this study is based on the Enron Email Dataset, originally obtained and made available through Kaggle. This source dataset consists of genuine emails collected from the Enron Corporation before it declared bankruptcy. The emails were part of an investigation by the Federal Energy Regulatory Commission. For the purposes of this project, we worked with a modified version of this dataset, specifically labeled and provided by the course instructors. This labeled dataset is derived from the original Enron Email Dataset and includes additional annotations necessary for our analysis of policy compliance and violation.

3.2 Size and Scope of the Dataset

The original dataset contains 517,401 emails, structured simply with two columns labeled 'file' and 'message'. For the purposes of this project, the dataset was reorganized to facilitate deeper analysis and more structured querying. The labeled dataset includes multiple columns: 'Date', 'From', 'To', 'X-To', 'X-From', 'X-cc', 'X-bcc', 'Subject', 'email_body', 'verdict', and 'violated_rules', which aid in specific analyses such as the detection of policy violations.

3.3 Characteristics of the Data

The dataset exhibits a wide range of communication types, reflective of a large corporate setting, encompassing everything from casual correspondence to formal reports and sensitive financial communications. A notable discovery in the labeled data was the presence of significant duplicates, nearly half of the dataset, which were not present in the original Kaggle release. The presence of duplicates in the labeled dataset, but not in the original or group-specific datasets, raised concerns about data integrity and processing errors, prompting thorough verification before their removal to ensure data quality.

Another point of interest is the handling of missing data. Specifically, the 'To' column exhibited null entries. It was assumed that these entries are not genuinely missing data, as the email body was often complete, suggesting that the email could have been a BCC or system-related message not directly addressed to visible recipients. This assumption guides the treatment of such entries in further analysis, avoiding unnecessary removal of data points.

3.4 Exploratory Data Analysis (EDA)

3.4.1 Purpose of EDA

The primary goal of Exploratory Data Analysis is to uncover insights without prior assumptions, identify any anomalies, and prepare the dataset for modeling. EDA is crucial as it informs the choice of analytical techniques based on initial findings, ensuring the data is well-understood and appropriately utilized for subsequent analysis.

3.4.2 Sequence and Techniques Employed in EDA

- **Duplicate Data Handling:** Initial analysis began with quantifying the number of duplicate rows in the dataset, followed by investigating their characteristics and removing them to ensure data quality and integrity.
- **Null Data Investigation:** The dataset was then examined for null data, particularly in the 'To' column, to determine the nature of these missing values and decide on their treatment, assuming non-impactful for the analysis given the presence of content in the email body.
- **Network Analysis:** Visualization of the communication network within the organization was conducted using network analysis plots. This helped identify the flow of information and key communicators within Enron.

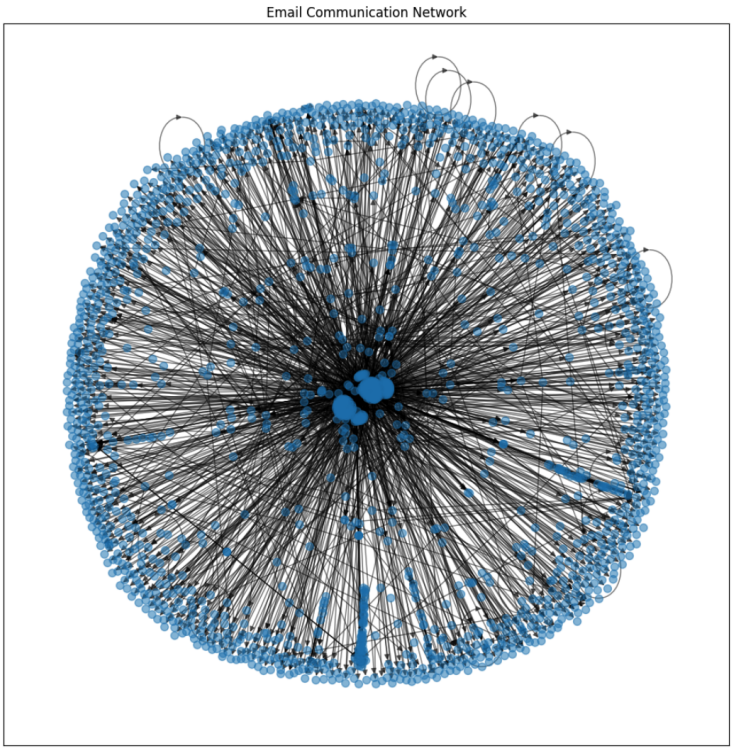
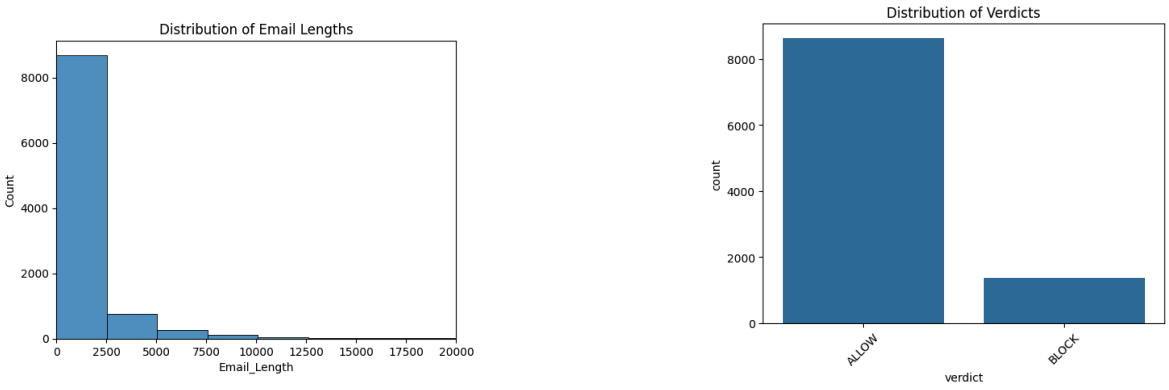


Figure 1: Network Analysis

- **Histograms of Email Characteristics:** Histograms were created to analyze the distribution of email lengths and verdicts (approved/blocked emails), providing insights into common communication patterns and compliance outcomes.

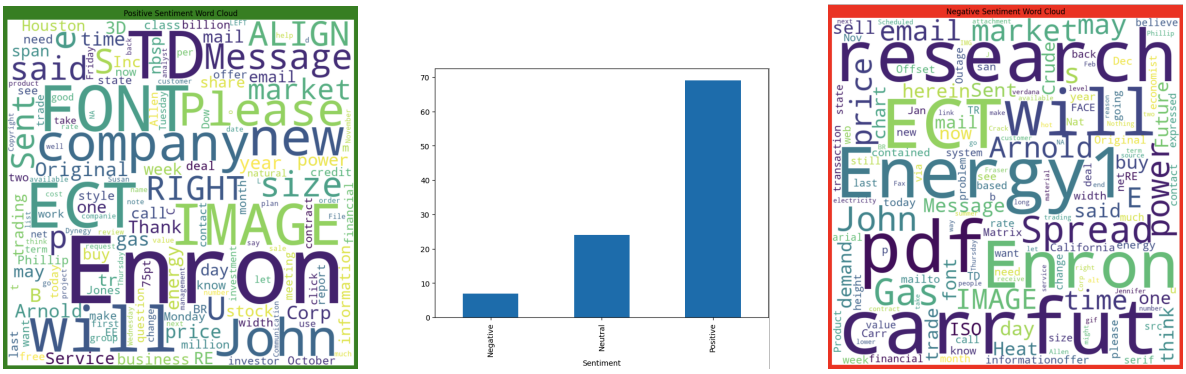


(a) Distribution of Email Lengths (b) Distribution of Verdict

Figure 2: Distribution of email lengths and verdicts

- **Sentiment Analysis:** The emotional tone of the emails was analyzed through sentiment analysis. This technique was instrumental in detecting underlying sentiments in communications, which could be indicative of compliance or ethical issues.

- **WordClouds and Sentiment Visualization Techniques:** To enhance the understanding of the dataset, various visual aids such as word clouds, sentiment plots, and bar plots were employed. A 'WordCloud' was particularly effective in highlighting common and frequent words, with 'stopwords' filtered out to emphasize terms specific to the corporate and email context, thus aiding in rapid thematic assessment. Additionally, the results from sentiment analysis were visualized through plots showing positive and negative sentiments. A bar plot categorizing words as positive, negative, or neutral further quantified and clarified the emotional tone across different communications within the dataset. This comprehensive visual representation was instrumental in assessing both the thematic and emotional dimensions of the communications.



(a) Positive Sentiment (b) Sentiment Bar Plot (c) Negative Sentiment

Figure 3: Sentiment Visualizations

3.4.3 LDA and Additional Visualization Techniques

The analysis involved using Latent Dirichlet Allocation (LDA) for topic modeling to uncover prevalent themes. To optimize the LDA, silhouette scoring assessed the separation of topics, determining the optimal number of topics. Subsequently, t-SNE visualization reduced the high-dimensional topic data to a two-dimensional space, allowing visual assessment of topic distribution and clustering. This approach confirmed the model's robustness by demonstrating clear separateness and coherence of the topics, providing deeper insights into the dataset's structure.

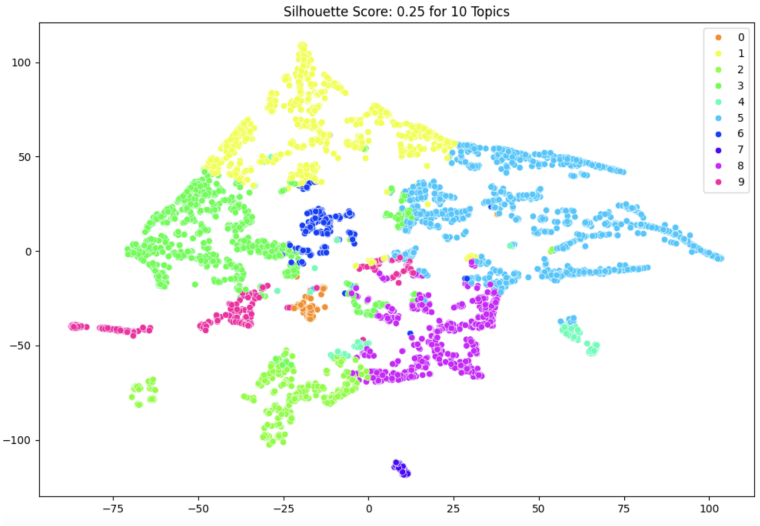


Figure 4: LDA and Additional Visualization Techniques

3.5 Additional Techniques for Policy Enforcement and Employee Identification

Techniques such as Named Entity Recognition (NER) for detecting personal and sensitive information, and checks for the sender's geographical location and corporate role were applied. These methods are essential for enforcing policies related to data privacy, geographical data transfer, and internal communication restrictions.

3.5.1 Policies Validation Section

The effectiveness of these techniques was tested against the dataset with ground truth labels to validate policy enforcement. The accuracy of these methods was quantified using the False Positive Rate (FPR) and False Negative Rate (FNR), crucial metrics for evaluating the effectiveness of the implemented policies.

4 Method

4.1 Methods used for content analysis - Rule-based + NER

1. **Rule-based approach (Regex):**

   - **Identity of email's sender and email's receivers:** We extracted data from the 'To' and 'From' columns, as they indicated which one has an Enron's employee email.

- **Affiliation of ECT or EES:** We extracted data from the 'X-To', 'X-From' and 'X-cc' columns, which usually indicated the affiliation of the department along with a "@" or "/" sign.
- **Sender/receivers geographic department (USA/EU):** We extracted data from the 'X-To', 'X-From' and 'X-cc' columns, which usually included the geographic department of the employee, the format of that information was usually the same with small minor changes, so we could construct a regex rule to extract this information and match the person to USA or EU.

4.1.1 Logic Behind Choices

- We chose regex as it is considered a relatively fast classification technique.
- Since all this data has a specific format, a regex would fit perfectly as it matches specific formats.

2. **Mixed approach of data learning and Rule-based (Regex):**

- **Identifying VP / Directors / C-Level employees:** We extracted specific data from our tagged data, as we know that rule 2.2 of our policy states that "sensitive business information can be passed only between VPs, Directors, and C-level employees", so, if we see a violation in our tagged data of rule "2.2", we know that the sender of that email is a VP / Director / C-Level employee, and we can create a list of those people. We also collected data from the internet (as Enron's company employees list is static and no longer changing), and we have added information from there as well.

4.1.2 Logic Behind Choices

- We have tagged data with the information that we look for, so we could use it to be more accurate in our prediction.
- We know that Enron's employees list is static as the company no longer exists, so we could use that information in order to look for the relevant people online.

3. **Named Entity Recognition (NER):**

- **Identifying PII and QID:** We have used the NER model to extract known labels that might indicate the presence of sensitive information in the email, our labels included phone numbers,social security numbers, location and more.

4.1.3 Logic Behind Choices

- Since PII and QID can come in many variations and dynamically change, we have preferred to use an ML model that is built to identify those traits.

4.2 Methods used for topic analysis - LDA

This method involves topic modeling using Latent Dirichlet Allocation (LDA) and visualization techniques.

1. **Topic Modeling:**

- **LDA Models:** Multiple LDA models with varying topic numbers are generated to identify latent topics within the email texts.

- **Dictionary and Corpus Creation:** Texts are tokenized and converted into a bag-of-words format suitable for LDA.

2. **Optimal Topic Selection:**
   - **t-SNE and Silhouette Analysis:** t-SNE reduces dimensionality for visualization, and silhouette scores evaluate the quality of clustering for each LDA model to select the optimal number of topics.
3. **Visualization and Interpretation:**
   - **Color Palette Generation:** A color palette is created for visualizing different topics.
   - **Document Formatting:** Emails are formatted with color-coded highlights to visualize the contribution of various topics.
   - **Topic Words Extraction:** Key words for each topic are extracted and listed for interpretation.

4.2.1 Logic Behind Choices

- **LDA for Topic Modeling:** LDA is effective in uncovering hidden topics in large text corpora and is suitable for the varied content of the Enron emails.
- **t-SNE and Silhouette Scores:** These techniques ensure the selection of the most coherent and distinct topic model, improving the interpretability of results.
- **Color Palette:** Enhances the visual differentiation of topics in formatted documents, making the analysis more intuitive.

4.2.2 Feature Selection and Parameter Tuning

Feature Selection:

- Processed Text: The primary feature for topic modeling, preprocessed to remove noise and standardize the text.
- Document Topics: LDA-derived topic distributions for each email.
- Keywords and Phrases: Identified using both LDA and predefined sets of finance and legal terms.

Parameter Tuning:

- Number of Topics: Evaluated using a range of values (e.g., 10, 15, 20, 25, 30) to find the optimal number based on silhouette scores.
- LDA Passes: Set to 10 for initial model training and 50 for the final model to ensure thorough topic discovery.
- t-SNE Parameters: Perplexity, learning rate, and iterations are tuned for effective dimensionality reduction and visualization.

4.3 Methods used for Finance Topics Identification

Topics are classified as finance-related using a predefined set of finance keywords and a similarity-based keyword expansion method. This process involves:

- **Keyword Set:** A base set of finance-related keywords is established.
- **Keyword Expansion:** Additional related terms are identified using word similarity measures.
- **Topic Classification:** Topics are analyzed to determine if they are finance-related based on the presence of these keywords.

4.4 Methods used for Legal Issues Identification

Emails are flagged if their subject or body text contains legal-related terms. This involves:

- **Regular Expression Search:** Identifies emails with legal-related content by searching for specific keywords in the subject and body text.
- **Flagging Legal Emails:** Emails that match the criteria are flagged as legal-related.

4.4.1 Logic Behind Choices

- **Keyword-Based Classification:** Provides a straightforward method to identify finance-related topics, leveraging domain-specific terminology.
- **Regular Expression Search:** Efficiently flags emails related to legal issues, focusing on key terms commonly associated with legal matters.

4.4.2 Feature Selection and Parameter Tuning

Feature Selection:

- Keywords: Both base and expanded sets for finance-related content.
- Regular Expressions: Specific terms related to legal issues.
- Processed Text: Preprocessed text of emails for analysis.

Parameter Tuning:

- Keyword Similarity Threshold: Set to a high value (e.g., 0.5) to ensure relevance.
- Regular Expression Patterns: Carefully crafted to accurately identify relevant emails.

4.5 Methods used for sentiment analysis

This method applies Natural Language Processing (NLP) techniques to analyze the Enron Email Dataset, focusing on sentiment analysis and feature extraction.

1. **Sentiment Analysis:**
   - **VADER Sentiment Analyzer:** Used to assess the sentiment of email texts.
   - **Sentiment Categorization:** Emails are classified as Positive, Negative, or Neutral based on the compound score.
2. **Feature Extraction:**
   - **Text Tokenization:** Breaks down email body text into individual tokens using RegexpTokenizer.
   - **Descriptive Statistics:** Computes text length, total word count, and vocabulary size.
   - **Word Frequency Analysis:** Identifies and prints the 50 most common words.
3. **Visualization:**
   - **Word Cloud Generation:** Creates visual representations of frequent words, filtered by sentiment.
   - **Sentiment Distribution Plot:** Displays a bar chart showing the distribution of sentiment categories.

4.5.1 Logic Behind Choices

- **VADER:** Chosen for its effectiveness in analyzing social media-like text, capturing both polarity and intensity of sentiment.
- **Text Tokenization and Statistics:** Essential for extracting meaningful features and understanding dataset characteristics.
- **Word Clouds:** Provide a visual summary of common terms, customized for clarity and relevance.
- **Sentiment Distribution Plot:** Offers a clear view of the sentiment landscape within the emails.

4.5.2 Feature Selection and Parameter Tuning

Feature Selection:

- Focuses on the 'email_body' text and derived sentiment scores.

Parameter Tuning:

- **Sentiment Thresholds:** Standard VADER ranges for sentiment classification.
- **Word Cloud Parameters:** Set for readability and relevance, with additional stopwords specific to the Enron dataset.
- **Visualization Settings:** Customized colors and sizes for clear and aesthetic presentations.

5 Results (on the labeled dataset)

5.1 Accuracy

- **False positive rate (FPR):** 0.15
- **False negative rate (FNR):** 0.35
- **True positive rate (TPR):** 0.64
- **True negative rate (TNR):** 0.84

While validating our model accuracy, we have noticed a high FNR. After investigation of the model's results, we have seen that this resulted from our lack of ability to identify rule '1.1' correctly. It seemed like we had not enough indications to accurately classify ECT/EES employees and 'legal' topics.

5.2 Performance

- **Resource usage:**

  Running the ML models

  - CPU usage: ∼99%
  - Memory usage: ∼30GB

We can see that our Machine Learning models consume a lot of resources. This is likely due to the mathematical calculations performed in our models (NER, LDA) and the size of the dataset, which contains over half a million samples.

Running on single email excluding ML models resource consumption

CPU usage: did not seem to change significantly.
Memory usage: did not seem to change significantly.

We see that aside from our heavy ML models, our other classifiers are very light and their resource consumption is very low.

- **Response time: It took us ∼10 milliseconds to classify each individual email in the dataset.** This runtime calculation does not include the runtime of the ML on the dataset.

6 Limitations

6.1 Current Limitations:

- **Scalability:** The current model has been tested on the static Enron email dataset and the labeled dataset, which are substantial but finite. Challenges may arise when scaling this model to continuously growing datasets or applying it in a real-time environment where emails must be processed and classified instantaneously. This could impact the practical deployment of our system in real-world scenarios where data volume and velocity are significantly higher.
- **Accuracy:** While the model performs well on historical data, issues such as false positives and false negatives persist, especially given the stringent thresholds set by the Corporate Security Officer (CSO). False positives, where non-violating communications are flagged, could disrupt normal business operations, while false negatives may lead to security breaches by missing crucial policy violations.
- **Performance Bottlenecks:** The complexity of the analyses, particularly with Latent Dirichlet Allocation (LDA) for topic modeling and NER for retrieving PII and QID identification, presents significant computational demands. These processes can be time-consuming and may not scale efficiently with larger datasets or in tighter real-time processing windows.

6.2 Future Work

1. **Methodological Improvements:**
    - **Advanced Machine Learning Algorithms:** Incorporating more sophisticated algorithms such as neural networks might improve the model's ability to generalize from the training data and reduce error rates, especially in distinguishing subtle nuances in email content.
    - **Deeper NLP Techniques:** Expanding the use of Natural Language Processing techniques could provide a deeper understanding of the content and intent of communications, potentially reducing misclassifications.
2. **Expansion of Data Sources:** To enhance the model's understanding and detection capabilities, integrating additional datasets, such as more recent corporate emails or datasets from different industries, could provide more diversity and volume, helping to refine the model's accuracy and adaptability to varied contexts.

3. **Real-World Testing:** Deploying the model within a controlled real-world environment, such as a pilot program within a corporation, would allow for the monitoring of the model's performance in real-time. This would not only test its effectiveness under actual operating conditions but also help in tuning the model in response to practical challenges and feedback.

4. **Continual Learning Framework:** Implementing a continual learning system where the model can adapt and update its parameters based on new data and emerging trends without needing retraining from scratch. This approach would help maintain its relevance and effectiveness over time, addressing changes in communication patterns and policy requirements.