**Introduction & Problem Formulation**

Security and privacy were not built into email when it was first invented, and despite email's importance as a communication method, these are still not built into email by default. As a result, email is a weak point for malicious or unintentional information leaks. [3]

**Problem Description: Developing a Mailing System Data Protection Solution**

Our objective is to create a robust data protection mechanism for a mailing system, enforcing predefined rules tailored to an organization's needs. In this task, we are tasked with constructing an email gateway capable of implementing six distinct roles. Leveraging a dataset comprising 517,401 emails from Enron Corp, the gateway must analyze each email, determining whether to allow or block it based on the predefined rules.

**Domain overview:**

A secure email gateway (SEG) is an email security product that uses signature analysis and machine learning to identify and block malicious emails or emails with content that not suppose to be send, before they reach recipients’ inboxes. They are important because email attacks, such as phishing and data leaks in emails are some of the most common cyber threats organizations face. SEG is intended to provide comprehensive protection against email-borne threats.

**How does an SEG work?**

An SEG functions by analyzing and screening email traffic to identify potential threats, utilizing signature analysis and machine learning algorithms.

Critical features of an SEG include [5]:

* Content Disarm and Reconstruction (CDR): Strips potentially malicious content from emails while maintaining their functionality.
* Sandboxing: Executes suspicious email attachments in a secure environment to assess their behavior.
* Data Loss Prevention (DLP): Prevents unauthorized transmission of sensitive information via emails.
* Anti-Phishing: Detects and blocks phishing attempts to prevent users from falling victim to fraudulent schemes.

**What Challenges does SEG has?**

* **Deployment and Configuration:**

Setting up a SEG requires careful configuration to ensure it effectively filters emails without blocking legitimate communication.

* **Accuracy and False Positives:**

SEGs must accurately identify and block malicious emails while avoiding false positives that could disrupt legitimate communication.

* **Performance and Scalability:**

SEGs must handle large volumes of email traffic efficiently without causing delays or bottlenecks in email delivery.

**Commonly used methods and our proposal:**

Our focus within our task and system will be on Data Loss Prevention (DLP), aiming to mitigate the risk of data leakage through email channels.

We plan to implement this using a combination of topic and content analysis methods, with further details to be provided later.

Creating an effective DLP policy is an ongoing process that involves multiple elements. During our research we have found out that there are various ways to approach DLP security and we’ve narrowed it down to five core components that serve as the backbone of any successful Data Loss Prevention strategy that we purpose to implement: Identification, Protection, Monitoring, Response, and Maintenance [2]

* **Identification**: Classify and locate sensitive information and patterns in the data.
* **Protection**: Implement measures like API gateway.
* **Monitoring**: Track data usage in real-time.
* **Response**: Act swiftly upon detecting unauthorized activities.
* **Maintenance**: Keep policies updated and compliant through regular reviews and staff training.

Through our prework research, we have gained insights into the importance of email security and the challenges associated with implementing effective data protection measures. Our proposed solution aims to address these challenges by leveraging advanced analysis techniques and comprehensive DLP policies.

**Related Work**

**Topic analysis:**

Topic analysis, as part of data loss prevention (DLP) strategies, is a machine learning technique used to automatically categorize text data. By analyzing unstructured text, such as emails and social media interactions, these tools can identify sensitive information that may violate DLP policies. This analysis helps businesses enforce their DLP policies by identifying and categorizing data that needs protection, such as personally identifiable information or confidential business data, allowing them to take appropriate actions to secure this information.[5]

We will elaborate about two techniques we learned about:

* **LDA -** One of the most common topic analysis techniques is Latent Dirichlet Allocation (LDA). **LDA** is a generative probabilistic model for collections of

discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. [6]

In other words, LDA is based on the assumption that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. This probabilistic model assigns topics to words and documents based on statistical patterns in the data.[7]

Using LDA for DLP policy enforcement can be instrumental in identifying and categorizing sensitive information within documents. By analyzing the topics identified by LDA, organizations can gain insights into the types of information being shared or discussed in their documents. This can help in enforcing DLP policies by flagging emails that contain sensitive information, such as legal or confidential business information.

* **Keywords analysis** - Using **keyword analysis** in the context of topic analysis is a text analysis technique that automatically extracts the most used and most important words and expressions from a text. It can help summarize the content of texts and recognize the main topics discussed.[8]

**Content analysis**

Content analysis is a research tool used to determine the presence of certain words, themes, or concepts within some given qualitative data (i.e. text). Using content analysis, researchers can quantify and analyze the presence, meanings, and relationships of such certain words, themes, or concepts.[9]

Two main techniques we have learned:

* **Dictionaries Based Approaches -** are commonly used in content analysis to classify texts into predefined categories based on the frequency of specific keywords. They can classify texts dichotomously or using scores, but their accuracy depends on how well the keywords align with the language used in the context. Using an 'off-the-shelf' dictionary outside its intended domain can lead to errors, as words may have different meanings in different contexts(*dictionaries may have difficulties in negation, irony, polysemy, and language-specific nuances*). Additionally, the writing conventions of different text types, like emails, can further complicate dictionary generalizability. While comprehensive keyword sets can yield reliable results, dictionaries are criticized for their simplicity and potential lack of completeness. To address these issues, word embeddings can be used to expand dictionaries and improve external generalizability and vocabulary coverage.[10][11]
* **Rule-Based -** usingregular expression (regex) in the context of content analysisis a sequence of characters that forms a search pattern. Regex are powerful patterns used for text matching and manipulation in computing. It can be used to perform operations like finding, matching, replacing, or splitting text based on that pattern**.**[12][13]

While powerful, regular expressions can become complex for advanced patterns. Care must be taken to properly escape special characters and understand regex behavior across different systems/languages. However, their flexibility and expressiveness make them an indispensable tool for robust text processing and content analysis.[8]

Using regex for DLP policy and email security, regular expressions can be leveraged to detect sensitive data patterns (credit cards, SSNs), identify restricted keywords/phrases, extract metadata like ages or genders from text, and more.

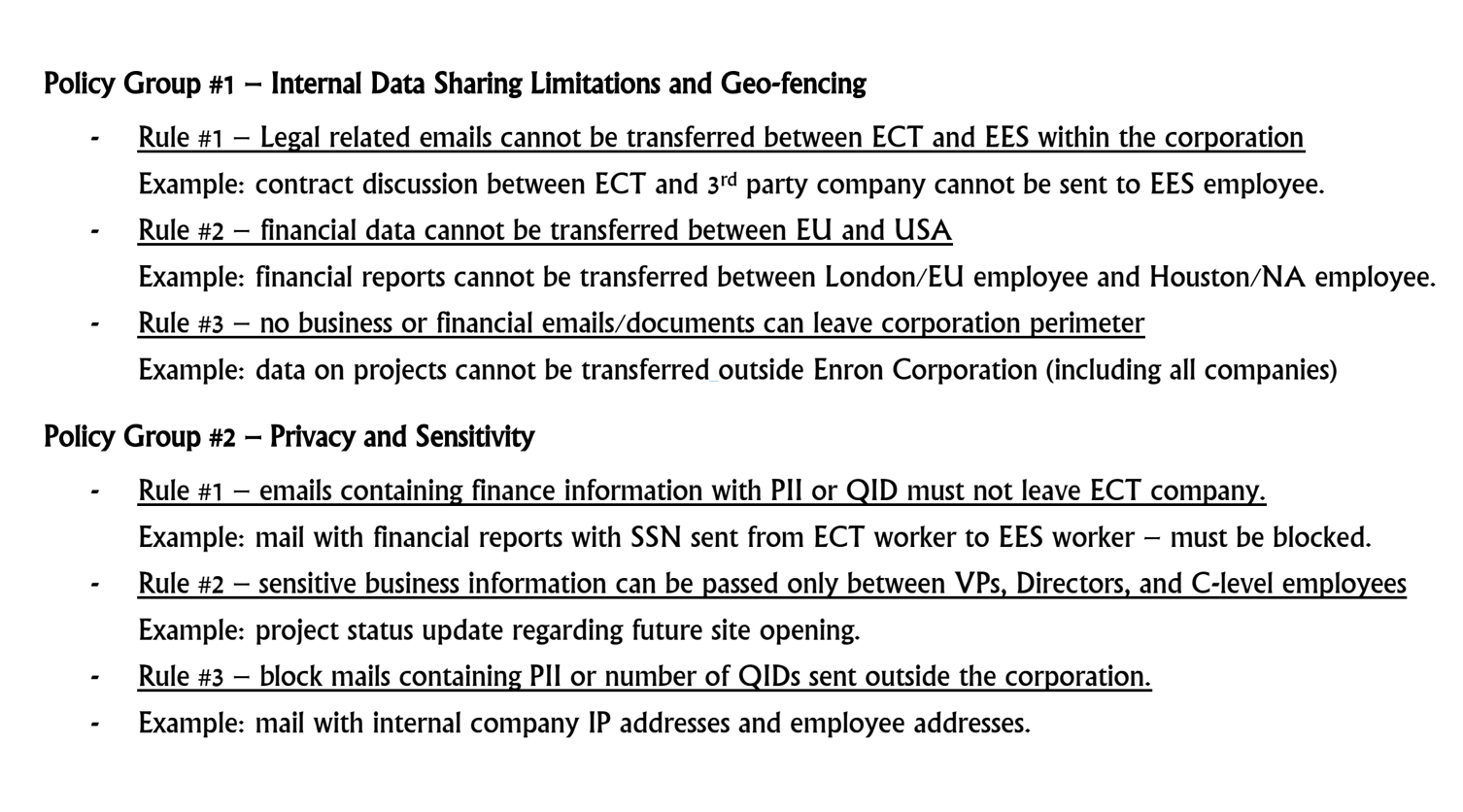
**Dataset Description**

The dataset utilized in our task comprises 517,401 authentic emails sourced from the Enron Corporation before its closure. Enron Corporation, headquartered in Houston, Texas, was an American company specializing in energy, commodities, and services.

On every email we have the following fields:

* **Date and Time**: record includes the date and time of transmission.
* **Sender Information**: The "From" field specifies the email address of the sender and the "X-From" may provide sender's name.
* **Recipient Information**: The "To" field indicates the recipient(s) of the email. The "X-To" may give their names.
* **Subject**: The subject of the email written by the sender.
* **Email Body**: The email content written by the sender.
* **Verdict and Violated Rules:** The "verdict" field specifies whether the email should be allowed or blocked based on the rules. (Indicate by the course team).

The "violated\_rules" list the broken rules.

 Our dataset also has the policy rules needs to be enforce:

**EDA**

During our EDA of a large volume of emails, we observed several patterns that can be leveraged to extract information about employees:

* **Forwarded Emails:** Forwarded emails often contain sender and recipient details, as well as information about the sender's department and location. These details typically follow a pattern like "Forwarded by [Employee Name]/[Location]/[Department]."
* **Training Data Insights:** The training dataset provides information about each email, indicating whether it was allowed or blocked, if blocked, which rules it violated. This data enables reverse engineering of rules and provides insights into employee details. We notice that some of the employers have the name of the company or department in their email address – direct indicate about the employee.
* **Multiple Recipients:** The "To" and "X-To" fields may contain multiple recipients, requiring careful consideration in rule checks and data extraction processes.
* **Company or Department Names in Email Addresses:** Some employees have the company or department name in their email addresses, providing direct indications of their affiliation
* **Patterns in Phone Numbers:** We identified specific patterns in phone numbers and adjusted our regex expressions accordingly for accurate extraction.
* **Enron-related Terms:** We identified common Enron-related words associated with the topics "legal", "financial" and "business" and incorporated them into our dictionary for enhanced data extraction.

Through our EDA, we uncovered these valuable insights that inform our approach to extracting employee information from emails.

**Method**

In our analysis, we utilized content analysis to extract Personally Identifiable Information (PII) and Qualified Identifiable Data (QID) from emails, along with metadata about the email itself and its sender and recipients.

For **content analysis**, we employed **Rule Engines** to identify and extract information such as phone numbers, Social Security Numbers (SSN), email addresses, zip codes, addresses, and credit card numbers. We implemented Regex expressions to define patterns for these common types of PII.

Additionally, we leveraged Spacy, a widely-used open-source NLP library in Python. Spacy provides efficient tools and algorithms for tasks like tokenization, part-of-speech tagging, named entity recognition, and dependency parsing. With its pre-trained models and broad language support, Spacy is instrumental in text analysis, information extraction, and various machine learning applications involving textual data.

To detect any potentially sensitive Personally Identifiable Information (PII) or Qualified Identifiable Data (QID) in emails, including instances that may not follow common patterns or are explicitly provided, we choose to used **Dictionary-Based Approache** to get those details. This involved constructing a dictionary containing sensitive words, which aided in identifying personal information within emails.

To conduct **topic analysis** and categorize the themes of emails, we employed **Keyword Analysis**. This entailed creating an extensive dictionary containing terms associated with topics such as "financial," "legal," and "business." By establishing a threshold, emails containing a significant number of words from these dictionaries were categorized under the respective topics.

**Structure** **and flow:**

We utilize the function "classify\_mail" to analyze an email and determine whether it should be blocked. This function incorporates both the previously mentioned content and topic analyses. It gathers data from these analyses and applies predefined rules to detect any violations.

**Learn from data:**

One of our primary mission involves learning from the provided data to classify employees into six groups: Enron employees, ECT employees, EES employees, USA employees, EU employees, and high-ranking employees. We gathered information on C-level executives and directors by conducting online searches[4].

We extract employee information, particularly from forwarded emails, predominantly in the format "Phillip K Allen/HOU/ECT," to identify employees affiliated with ECT or EES and their respective locations.

Additionally, from emails that violate rules, such as Rule #3 group 2, which pertains to blocking emails containing PII or QIDs sent outside the corporation, we infer sender and recipient information. For instance, if an email violates this rule, indicating the sender is from Enron, we learn about the employees involved.

We continually learn from the entire training dataset to maintain these six sets of employees.

**Limitations and Future Work**

**//הקדמה**

One limitation is the reliance on patterns in Rules Engine content analysis and rules for data extraction, which may not capture all variations and nuances present in real-world email data.

Future work could explore more sophisticated machine learning techniques to enhance the accuracy and flexibility of information extraction.

Additionally, our current focus on specific types of information, such as employee names and departments, may overlook other relevant details that could provide valuable insights. Future research could expand the scope of data extraction to include a broader range of employee attributes and contextual information.

Furthermore, our analysis is based on a specific dataset from Enron Corporation, which may not fully represent the diversity of email communication patterns and organizational structures found in other contexts. Future studies could validate our approach on a more diverse set of datasets to assess its generalizability and effectiveness across different organizations and industries.

Additional limitation of our email analysis system and an important aspect to address is the handling of new senders for whom we lack detailed information. Currently, our system relies on predefined rules and patterns to extract employee details from emails. However, when encountering emails from new senders not present in our existing database, the system may struggle to accurately identify and classify them.

To mitigate this limitation, one approach is to expand the repertoire of methods used for analyzing incoming emails, reducing reliance on existing data.

REFERENCES

[1] Check Point article about SEG - <https://www.checkpoint.com/cyber-hub/threat-prevention/what-is-email-security/what-is-a-secure-email-gateway-seg>

[2] Zoran Cocoara , October 20, 2023, DLP Policy 101: From Basics to Best Practices, https://www.endpointprotector.com/blog/dlp-policy-101/

[3] Cloudfare article , <https://www.cloudflare.com/learning/email-security/what-is-email-security/>

[4] From Wikipedia, the free encyclopedia ,https://en.wikipedia.org/wiki/Enron

[5] <https://monkeylearn.com/blog/what-is-topic-analysis/>

[6] https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[7] <https://monkeylearn.com/blog/introduction-to-topic-modeling/>

[8] https://monkeylearn.com/keyword-extraction/

[9] https://www.publichealth.columbia.edu/research/population-health-methods/content-analysis

[10]

[https://link.springer.com/chapter/10.1007/978-3-658-36179-2\_3](https://link.springer.com/chapter/10.1007/978-3-658-36179-2_3 [11])

[[11]](https://link.springer.com/chapter/10.1007/978-3-658-36179-2_3 [11])

[https://www.tandfonline.com/doi/full/10.1080/2474736X.2022.2029217](https://www.tandfonline.com/doi/full/10.1080/2474736X.2022.2029217 [12])

[[12]](https://www.tandfonline.com/doi/full/10.1080/2474736X.2022.2029217 [12])<https://andy.egge.rs/teaching/content_analysis/content_analysis_week_1_2016_hw.pdf>

[13] <https://towardsdatascience.com/the-utility-of-regular-expressions-in-data-science-babf94285966>