1. Find a larger and more recent dataset.

**Ans:** I found a few datasets related to email phishing. I attached the dataset file with this document that included the description of each document. Please read the document and let me know which dataset that you feel is more suitable for your proposal. I will update your proposal accordingly.

1. Why are we using TF-IDF and Count Vectorizer in our feature extraction methods? Send me some material about it.

**Ans:** TF-IDF (Term Frequency-Inverse Document Frequency) and Count Vectorizer are commonly used feature extraction methods in natural language processing (NLP). These methods help transform text data into numerical representations that can be utilized by machine learning algorithms.

* Count Vectorizer: Count Vectorizer is a simple method that converts a collection of text documents into a matrix of token (word) counts. It builds a vocabulary of all the unique words in the documents and assigns a numerical value to each word based on its frequency of occurrence in each document. The resulting matrix represents the occurrence of words in the documents. Count Vectorizer is a useful technique for text classification tasks where the frequency of occurrence of words is significant. However, it does not consider the relative importance of words in the corpus.
* TF-IDF: TF-IDF is a more advanced technique that takes into account both the term frequency (TF) and the inverse document frequency (IDF). TF represents the frequency of a term (word) within a specific document, while IDF measures the rarity of a term across the entire corpus. The TF-IDF score is calculated by multiplying TF and IDF. TF-IDF helps to identify the importance of words in a document relative to the entire corpus. Words that appear frequently in a specific document but rarely in others are considered more significant.

By using TF-IDF or Count Vectorizer, we can convert text data into numerical features that machine learning models can understand and analyze. These methods enable us to capture important information from text documents and use it for various NLP tasks like sentiment analysis, text classification, and information retrieval.

Here are some recommended online resources to learn more about TF-IDF and Count Vectorizer:

* Scikit-learn documentation:
  + TF-IDF: https://scikit-lear n.org/stable/modules/generated/sklearn.feature\_extraction.text.TfidfVectorizer.html
  + Count Vectorizer: <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>
* Natural Language Processing with Python book:
  + Chapter 6: Learning to Classify Text - <https://www.nltk.org/book/ch06.html>
* Machine Learning Mastery:
  + TF-IDF tutorial: <https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/>
  + Count Vectorizer tutorial: <https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/>

These resources should provide you with a solid foundation and practical examples to understand and implement TF-IDF and Count Vectorizer for feature extraction in NLP.

1. What are the features that will be extracted from the email body, header, and attached URLs?

**Ans:** We will extract the lexicon features from the email body, header and attached URLs.

1. Is the proposed model unique? and why?

**Ans:** The proposed model is not unique in its nature, as the ensemble learning technique has been used in numerous studies. However, the combination of different models for predicting the label of email.

1. What are the reasons for using four different models (M1, M2, M3, and M4)?

**Ans:** The aim of the proposed work is to detect phishing in emails by considering the email header, body and attached URLs. In the proposed work, the features will be extracted from the text of each part. If the extracted features will pass to single model the it may not be able to extract the pattern of phishing email in all sections of email. TO overcome this problem, the extracted features from text will pass to individual model that will learn the pattern. Lastly, the values from 2nd last layer of each model will be integrated and pass to the finalized model.

6. What about deep learning for Anomaly Detection?

Sequence-to-sequence models, like we can use an encoder-decoder structure for Anomaly Detection.

What do you think?

**Ans:** Not understandable. Please explain.