Notes

We're observing strong model performance from a simplistic approach due to inherent patterns within the data. These patterns suggest that emails from specific users or with similar subjects often fall into the spam category. Consequently, by conducting minimal feature engineering and training a basic model on this data, we achieve high accuracy, precision, and recall, with a focus on precision.

However, to develop a more robust model capable of real-world deployment, we need access to additional diverse datasets. These datasets should encompass a broader range of trends and patterns, facilitating the creation of a more comprehensive and adaptable model. By incorporating such diverse data, our model can better generalize to unseen scenarios and effectively handle the complexities inherent in real-world email environments.

Also other complicated models (Neural networks, pre trained & fine tuned LLMS) were not tried due to time constraint as well as no requirement given the data quality and size

Area of Improvement

1. Create more meaningful features
2. Use different types of embedding Glove, Word2Vec, FastText
3. Experiment with contextual based embeddings based on Transformers (BERT)
4. Use a pretrained GPT model (preferably LLama2)
5. Fine tune the model using annotation
6. Experiment with alternative architectures such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), or hybrid models combining CNNs and RNNs
7. Hyperparameter tuning
8. Data augmentation
9. Feed in External data for spam & noise detection