Spam Detection: Traditional Machine Learning and Deep Learning Approaches

Spam detection remains one of the biggest challenges in email security, with traditional machine learning and deep learning having their share of advantages and disadvantages. Traditional approaches would generally rely on pre-processed feature engineering using algorithms such as Naïve Bayes, SVM, and Random Forest classifiers, with features including occurrences of suspicious keywords, sender reputation, and aspects of email headers. These approaches constitute lightweight computations that are more interpretable; thus, they can be used for first-level spam filtering in scenarios such as resource-constrained environments. However, such methods would fare poorly against sophisticated spam techniques including adversarial attacks and contextually intricately-crafted phishing attempts.

On the contrary, RNN, CNN, and Transformer-based deep learning models (e.g., BERT) can work completely end-to-end to extract features directly from raw email content. Deep Learning systems detect subtle patterns that would allow the confusion of text-based spam, image-based spam, and semantically complex phishing emails with very high accuracy and adaptability, which require heavy computational power for examples GPUs for training and large amounts of labelled data. But all this makes them stand in a black box, with the inability to explain why a certain decision was made, which is sometimes a drawback in regulated domains.