Email Fraud Detection

Pooja Shinde

DeVos Graduate School, Northwood University MIS 661: Artificial Intelligence & Business Analytics

Dr. Mighty Itauma

March 4th, 2025

Model Performance Comparison:

I've used the codes for multiple models such as Logistic Regression for email fraud detection and Conditional Random Fields for sequence labelling. Here's a summary of the metrics for these models:

Logistic Regression Model:

Precision: It measures the quantity of useful recognition which is literally correct. Fewer false precision means few false positives.

Recall: It measures the quantity of actual positives that were exactly identified. Fewer false negatives mean high recall.

F1-Score: Harmonic mean of precision and recall, gives a balance between the two metrics.

ROC-AUC: It explains how effectively the model differentiates between classes. A higher score means better performance.

The Logistic regression model wholly depends on precision, recall, F1-score, and ROC-AUC. Additionally, it works skilfully for classification tasks where you have enough data to train the model.

The **Logistic Regression** model is evaluated based on its precision, recall, F1-score, and ROC-AUC. Typically, it works well for classification tasks where you have enough labelled data to train the model. The model assumes that there is a way to clearly distinguish fraudulent emails from non-fraudulent ones based on the features in the email.TF-IDF helps to convert the text data into numerical features. It helps to show the importance of words in the emails or content to find the difference between fraudulent and non-fraudulent emails.

Sequence Labelling using CRF:

This model is used for sequence labelling tasks like Named Entity Recognition which is a key for analyzing sequences like email content, where you may need to recognize the specific named entities like departments or people.

The CRF model aligns each token which will consider the context around each word to predict the label. It does this only after making the predictions based on the sequences structure, to make it useful for tasks where the order of tokens is equally important.

Why One Model might perform Better:

Logistic regression works better when you deal with simpler classification tasks that depends on fixed features such as specific word or patterns. Hence, the model TF-IDF features represents the importance of words and not the sequence.

CRF focuses more on the sequence or order of data such as sequence labelling. It is useful to target the specific email components like names, locations, and organizations which will help to detect the fraud if there is possibility of fraud indicators.

Assumptions of Each Model:

Logistic Regression: It represents the straight-line relationship between the words of an email and whether the email is fraud or not. It tries to find the outcome based on the characteristics of the content. It doesn't consider the full meaning or contexts together; it analyses it one by one. Also, it shows that each word in the email is independent from other words.

CRF: CRF looks at the entire sequence of words and uses previous words to help predict the label of the current word, which makes it great for tasks involving sequences of data (like text).

Importance:

In Logistic Regression, specific words or phrases that are frequently found in fraudulent emails is of high importance. Trained Logistic Regression model helps to extract features using the coef_ methods.

In CRF, it's a full concept of sequence-based model and not feature-based model. But you can still analyse the transitions between labels such as how often an 'o' token follows a B-ORG token to gather which token or sequence of tokens are important.

Business Context:

For fraud detection in emails, the business contexts would suggest using a model that will identifies some risky keywords such as "urgent transfer" or the existence of certain organizations. If the main goal is to classify emails as fraud or not, Logistic regression is better to handle the large datasets with high-capacity features.

If the task is more complex and involves detecting fraud content based on the connection between entities such as identifying a suspicious email based on the sequence of features like "urgent" followed by a "bank "mention, a CRF model is better in providing the performance. Sequence-based models are excellent for tasks where sequence and order matters.

Conclusion:

Logistic Regression works better when its detection relies on certain keywords or features.

CRF performs better if the sequence is difficult or looking for patterns that indicate how tokens are inter-connected to each other.

In practice, both models could be useful depending on the nature of the problem. Logistic Regression is a strong baseline for simpler tasks, whereas CRF can help in more nuanced problems where the context between words or entities is critical.