SFWRENG 4NL3
Project Step 3 - Annotation Task Report

Sumanya Gulati

March 2025

1 Description of Task

The task consisted of going through emails and categorizing them as potential phishing emails or genuine emails.

2 Interesting Aspects

As someone who grew up in the age of technology, I have always believed that I am tech-savvy enough to not fall for phishing scams but for a lot of the advertising or marketing emails, it took me a bit to be sure about my answer. For a bunch of them, I could not be 100% certain about whether it was a genuine marketing campaign or a phishing scam.

3 Insights and Challenges

Based on the annotation guidelines, for a lot of the emails that had gibberish in them and seemed incomprehensible to me, I had to classify them as *not a phishing scam* solely because they did not contain a URL.

As for the model, I believe that there might be discrepancies in the annotator agreement ratio pertaining to the classification of marketing emails, order and shipping confirmation emails and such. Based on how elaborately the model is trained and how robust the classification criteria is for such emails, the model might struggle to accurately label these emails.

4 Mental Model

For every email that either started with a 'Dear ¡name¿' or ended with a 'Thanks ¡name¿', I could automatically classify them as *not phishing* saving me from having to read the entire email. Based on the guidelines, for emails that were too long (and there was a substantial number of them), I skimmed across the content and looked for any URLs. If it did not have a URL, I spared myself the effort of having to read all those sentences and labelled it as *not phishing*. I followed the same principle for emails that were either forwarded or started with 're' because it was evident that those ones were written by a colleague.

As for the marketing campaigns and order/shipping confirmation emails, if an email had wonky punctuation or just seemed sketchy, I classified it as a *phishing scam*. Also, as outlined in the guidelines, whenever in doubt, I labelled the email as a *phishing scam*.

5 Ambiguity in Instructions

The instructions were pretty clear and the guideline to classify any emails that could be determined with a 100% certainity as a *phishing email* simplified matters a lot because I ran into this dilemma very often. The only suggestion I have is to perhaps include information about system generated emails such as the ones about 'could not deliver email because the address is invalid' or 'no variance detected' and more. These emails did not seem suspicious enough to be phishing emails and were probably system generated emails with company-relevant information.