Classifiers Performance Report (Prelim)

Business goal: identify a minimal possible set of emails that include PII.

Requirements:

- 50k sample set should run in <15 mins.
- The goal is to create a workflow that is cheap, accurate, and fast (in priority order).

Note: Using the gold labelled subset, most emails don't have PII.

Solution: Have classifier that has low false negative rate (to not miss any PII emails) and with high true negative rate (to cut down the number of emails for manual review).

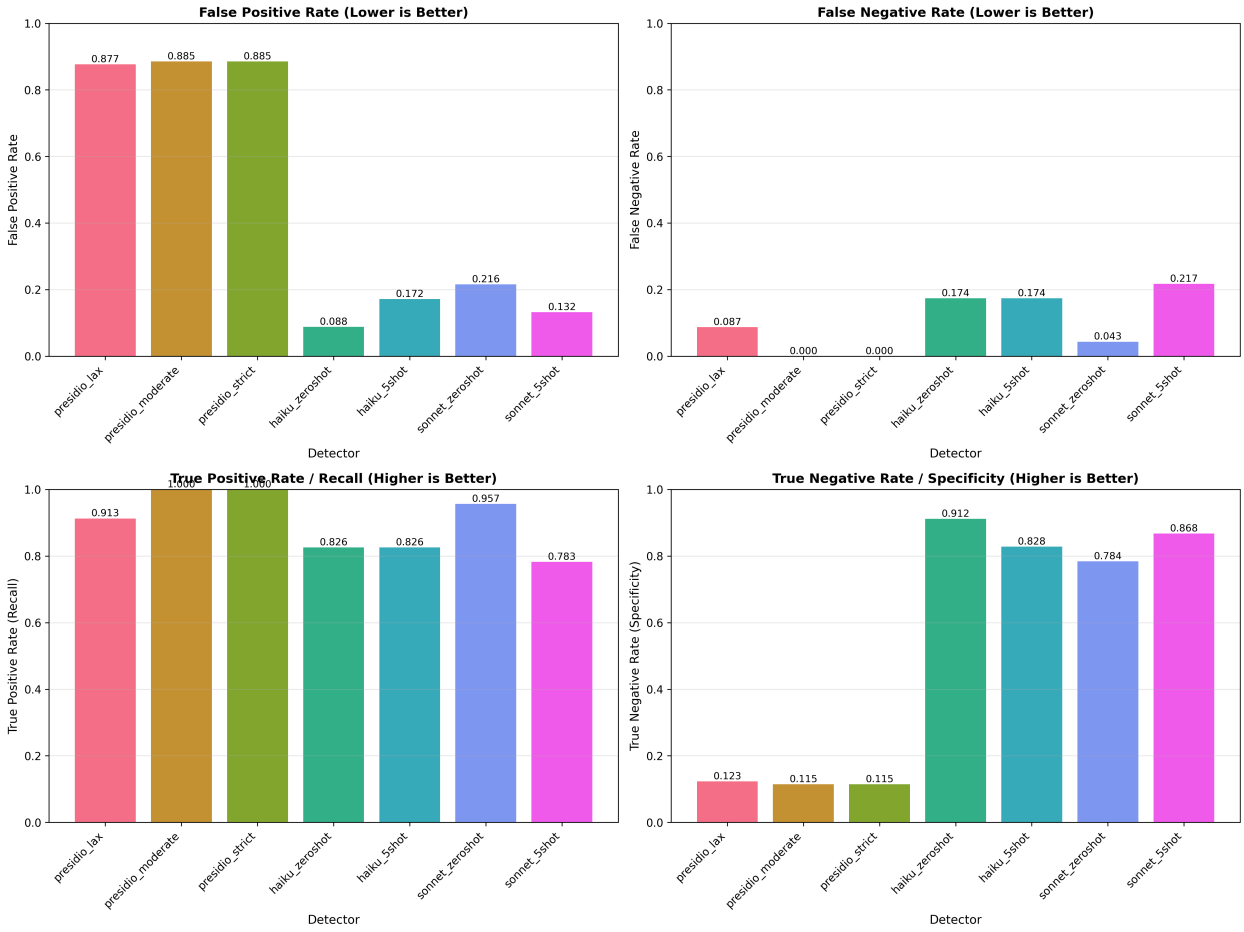
Methods

We tested out seven classifiers with different accuracies, cost and speed.

- Presidio (lax), uses Microsoft's Presidio analyzer. PII is only classified if it scores confidence above 0.8.
- Presidio (moderate), uses Microsoft's Presidio analyzer. PII is only classified if it scores confidence above 0.5.
- Presidio (strict), uses Microsoft's Presidio analyzer. PII is only classified if it scores confidence above 0.3.
- 4.5 Haiku zero-shot (with a prompt explaining the PII specifications)
- 4.5 Haiku 5-shot (with a prompt explaining the PII specifications, and five examples)
- 4.5 Sonnet zero-shot (with a prompt explaining the PII specifications)
- 4.5 Sonnet 5-shot (with a prompt explaining the PII specifications, and five examples)

Results

- Presidio (moderate and strict) satistifies FNR=0, but TNR is only at ~10%.
- Sonnet 4.5 zero shot has the next lowest FNR at 4.3%, and a good TNR at 78%. However, the non-zero FNR is not desirable.



Next steps

- Research on how to reduce TNR of presidio while maintaining 0 FNR
- Research on how to reduce FNR of LLM-based classifiers to 0
- Currently, Presidio is projected to take 20 minutes to run the full 50k dataset. Research on how to make this faster e.g. host an EC2 machine.