Ivan Miller. Reading Summary "Exploiting Machine Learning to Subvert Your Spam Filter". 11.22.2022

1. Explain the main assumptions the authors make about adversary capabilities to enable their attacks.

The main assumption is that the attacker is capable of contaminating the training set by sending emails that will be labeled as spam during retraining of the model. Generally an attacker may have one of two main goals:

1) **Expose the victim to a spam email** - which may be achieved by executing a so-called *dictionary attack* and sending emails that contain many words that are likely to occur in a normal non-spam email. After retraining the model on data that includes these emails, words from the attack emails will be receiving higher spam score, which could increase the chances of ham email being classified as spam (if it contains words from the attack email).
2) **Prevent the victim from receiving (seeing) a legitimate (ham) email** - which is done by sending the victim emails containing words likely to occur in the target email. When the model is trained on the data that includes the attack email, the spam scores of the targeted words increase, and the target message is more likely to be classified as spam.

2.Explain one of the proposed defenses against the attacks.

According to the paper, *Reject On Negative Impact (RONI)* defense is very successful against dictionary attacks, since it can identify all attack emails without falsely flagging any non-attack emails. Since each attack email increases the likelihood of ham messages being incorrectly filtered, the main idea behind that defense is to measure the impact of each email by testing the difference in performance with and without that email, having a goal of removing malicious emails from the training data.

The process is roughly as follows: one by one each email (Q) is being set aside from the training set (T) and the model is then trained on both T and (T - Q) datasets separately.
After that the performance of both models is being tested on a validation set, and by recording a change in incorrect classifications between T and T-Q models we could measure the impact of an individual email Q on the result. In an event that a given email Q has a significantly negative impact, that email is being removed from the training set.

3.Can you think of any new proposed defenses that aren't covered here? Or can you think of any additional avenues of attack that aren't explored?

I wonder if it would be possible to detect instances of negative influence on a model after retraining by tracking any increases in misclassifications using the same validation set and two states of the model: before and after retraining.

4. What's a further question you would like to see answered in follow-up work?

I thought that the idea behind the dynamic threshold defense was clever but I did not understand the logic behind the choice of thresholds. It appears that the values selected for threshold Θ_0 for ham were lower than the one for the SpamBayes (0.05 and 0.10 vs. 0.15) and higher for spam threshold Θ_1 (0.95 and 0.90 vs 0.90) for the Spam Bayes, so I would've loved to learn more about the logic behind those choices.