Abstract Using statistical machine learning for making security decisions introduces new vulnerabilities in large scale systems. We show how an adversary can exploit statistical machine learning, as used in the SpamBayes spam filter, to render it useless—even if the adversary’s access is limited to only 1% of the spam training messages. We demonstrate three new attacks that successfully make the filter unusable, prevent victims from receiving specific email messages, and cause spam emails to arrive in the victim’s inbox.

2.1 Introduction

Applications use statistical machine learning to perform a growing number of critical tasks in virtually all areas of computing. The key strength of machine learning is adaptability; however, this can become a weakness when an adversary manipulates the learner's environment. With the continual growth of malicious activity and electronic crime, the increasingly broad adoption of learning makes assessing the vulnerability of learning systems to attack an essential problem.

The question of robust decision making in systems that rely on machine learning is of interest in its own right. But for security practitioners, it is especially important, as a wide swath of security-sensitive applications build on machine learning technology, including intrusion detection systems, virus and worm detection systems, and spam filters [13, 14, 18, 20, 24].

Past machine learning research has often proceeded under the assumption that learning systems are provided with training data drawn from a natural distribution of inputs. However, in many real applications an attacker might have the ability to provide a machine learning system with maliciously chosen inputs that cause the system to infer poor classification rules. In the spam domain, for example, the adversary can send carefully crafted spam messages...
that a human user will correctly identify and mark as spam, but which can influence the underlying machine learning system and adversely affect its ability to correctly classify future messages.

We demonstrate how attackers can exploit machine learning to subvert the SpamBayes statistical spam filter. Our attack strategies exhibit two key differences from previous work: traditional attacks modify attack instances to evade a filter, whereas our attacks interfere with the training process of the learning algorithm and modify the filter itself; and rather than focusing only on placing spam emails in the victim's inbox, we also present attacks that remove legitimate emails from the inbox.

We consider attackers with one of two goals: expose the victim to an advertisement or prevent the victim from seeing a legitimate message. Potential revenue gain for a spammer drives the first goal, while the second goal is motivated, for example, by an organization competing for a contract that wants to prevent competing bids from reaching their intended recipient.

An attacker may have detailed knowledge of a specific email the victim is likely to receive in the future, or the attacker may know particular words or general information about the victim's word distribution. In many cases, the attacker may know nothing beyond which language the emails are likely to use.

When an attacker wants the victim to see spam emails, a broad dictionary attack can render the spam filter unusable, causing the victim to disable the filter (Section 2.3.1.1). With more information about the email distribution, the attacker can select a smaller dictionary of high-value features that are still effective. When an attacker wants to prevent a victim from seeing particular emails and has some information about those emails, the attacker can target them with a focused attack (Section 2.3.1.2). Furthermore, if an attacker can send email messages that the user will train as non-spam, a pseudospam attack can cause the filter to accept spam messages into the user's inbox (Section 2.3.2).

We demonstrate the potency of these attacks and present a potential defense—the Reject On Negative Impact (RONI) defense tests the impact of each email on training and doesn't train on messages that have a large negative impact. We show that this defense is effective in preventing some attacks from succeeding.

Our attacks target the learning algorithm used by several spam filters, including SpamBayes (spambayes.sourceforge.net), a similar spam filter called BogoFilter (bogofilter.sourceforge.net), the spam filter in Mozilla's Thunderbird (mozilla.org), and the machine learning component of SpamAssassin (spamassassin.apache.org)—the primary difference between the learning elements of these three filters is in their tokenization methods. We target SpamBayes because it uses a pure machine learning method, it is familiar to the academic community [17], and it is popular with over 700,000 downloads. Although we specifically attack SpamBayes, the widespread use of its statisti-