Spam or Ham: Detecting Spam Texts

DS 4002 Case Study by Jessica Li

Spam texts are becoming increasingly prevalent and harmful in the daily lives of people across all demographics. In the first half of 2023, over 78 billion spam texts were received by Americans alone. These spam texts caused an estimated loss of $13 billion within that timeframe, and these damages are only increasing in number and magnitude as technology continues to evolve.

Spam texts are unwanted and unsolicited texts sent to a large number of recipients. They are sent for various reasons, including commercial purposes, fraudulent purposes, malware, or viruses. Spam texts can be sent by real humans or, more commonly, botnets.

You are a data scientist tasked with determining whether or not certain keywords are associated with spam texts. In order to detect spam text, you will use text analysis techniques to identify if text messages containing a higher frequency of specific keywords are more likely to be classified as spam compared to non-spam messages. All the materials you need will be provided in this GitHub repo: https://github.com/aqj6td/DS4002_CS3/tree/main. You will first check through the dataset to scan for duplicate and stop words. Next, you will calculate the probability of each word appearing in spam and non-spam emails. Then, you will create metrics to test against your text analysis model.

This task is important and has vast implications. Identifying keywords for spam texts could impact the lives of many, helping people of all ages protect themselves against this increase in social engineering hacking.

"Spam Texts Trends | RoboKiller," www.robokiller.com. https://www.robokiller.com/spam-text-insights