Email Spam Detection Case Study

By: Somphors Yun

Table of Contents

1. Problem Statement

Image Source

What is an Email Spam?

Spam emails are **unsolicited** email messages that are sent by **people you don’t know**. They are almost always **commercial** and driven by a **financial motive**.

It may includes:

- Promotional emails that you did not ask for
- Counterfeit messages that attempt to trick you into giving out personal information
- Fraudulent messages from hacked email accounts

Did You Know?

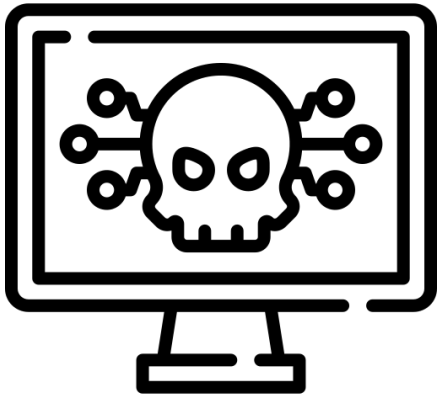
May 2019

Spam emails constituted almost **85%** of total emails sent globally

Why is Spam a Problem?

Waste of Time

Means for Malwares



Exploit Data Privacy

Why Do We Need Email Spam Detection?

Shielded From Attacks

- Prevent SPAM from getting into the inbox → spam filtering
- Effective anti-malware tools
- Reduce the risk of users clicking on things they shouldn't

Better Quality of Life

- Run smoother
- Used for desired purpose only
- Save time

2. About the Dataset

Image Source

Dataset of Email Spam

- Contain the information of 5728 mails
- Dataset is categorical → Classification
- Two columns:
  - Text: context of mail
  - Spam:
    - 0 → NOT SPAM
    - 1 → SPAM

3. Algorithm Comparisons

Naive Bayes

- **Supervised learning** which based on **Bayes Theorem** and used for **classification** problems
- Mainly used in **text classification** that include **high-dimensional** dataset
- One of the **simple** and **most effective** classification algorithms
- Some popular examples are: spam filtration, sentimental analysis, and classifying articles

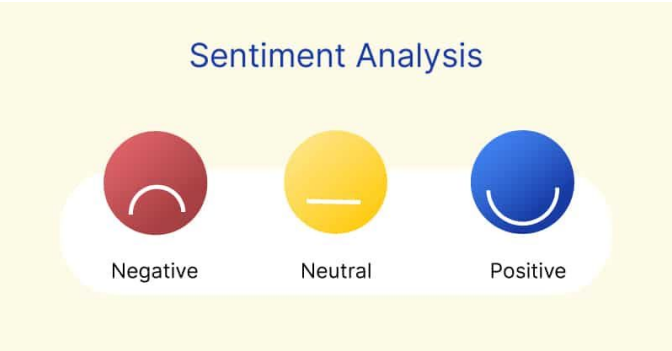


Image Source

Support Vector Machine

- **Supervised learning** aim to create the **best line** or **decision boundary** that can **segregate** the data into classes
- Used for both **classification** and **regression** problems
- Suited for **linear** and **non-linear** dataset
- **Faster** prediction along with **better accuracy** compared to other classification algorithms

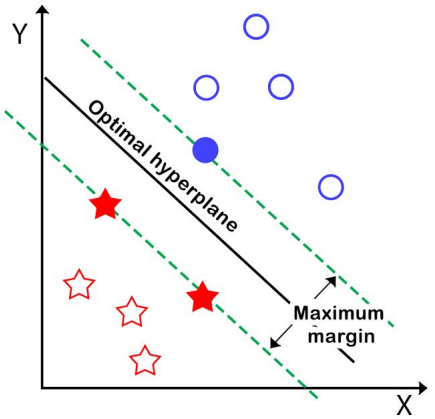


Image Source

Comparisons

Naive Bayes

- Supervised learning
- Linear classifier
- Classification
- Fast & easy algorithm to predict a class of dataset
- Perform well in multi-classes classification

SVM

- Supervised learning
- Linear & non-linear classifier
- Classification & regression
- Effective in high dimensional spaces
- Doesn't support multi-classes classification
- Faster prediction with high accuracy

4. Code Demonstration

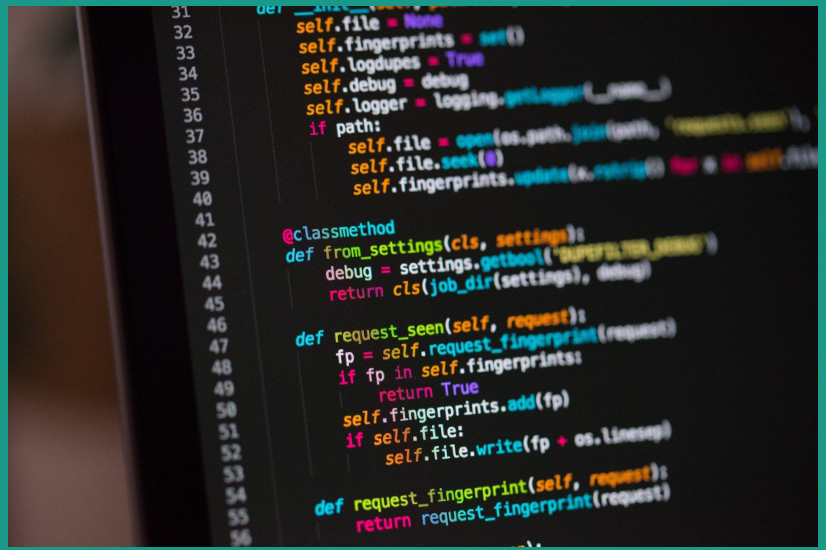


Image Source

References

- https://www.sciencedirect.com/science/article/pii/S2405844018353404#:~:text=Some%20of%20the%20most%20popular,them%20used%20RBFNN%20for%20classification.
- https://www.hindawi.com/journals/scn/2022/1862888/
- https://www.youtube.com/watch?v=cNLPt02RwF0&ab_channel=ComputerScience
- https://thecleverprogrammer.com/2021/06/27/spam-detection-with-machine-learning/
- https://thecleverprogrammer.com/2021/07/06/end-to-end-spam-detection-with-python/
- https://www.youtube.com/watch?v=exHwwy9kVcg&ab_channel=CodeHeroku
- https://towardsdatascience.com/spam-detection-in-emails-de0398ea3b48
- https://www.youtube.com/watch?v=CKs_vxCqMlw&ab_channel=LearnwithIntellify
- https://becominghuman.ai/spam-mail-detection-using-support-vector-machine-cdb57b0d62a8
- https://blog.fivenines.com/spam-filtering-why-its-important-and-how-it-works#:~:text=Implementing%20spam%20filtering%20is%20extremely,used%20for%20their%20desired%20purpose.
- https://www.geeksforgeeks.org/naive-bayes-classifiers/#:~:text=Naive%20Bayes%20classifiers%20are%20a,is%20independent%20of%20each%20other.
- https://www.javatpoint.com/machine-learning-naive-bayes-classifier
- https://towardsdatascience.com/na%C3%AFve-bayes-spam-filter-from-scratch-12970ad3dae7#:~:text=Spam%20Filtering,-Photo%20by%20Hannes&text=With%20Bayes'%20Rule%2C%20we%20want,to%20be%20used%20in%20classification.
- https://scikit-learn.org/stable/modules/naive_bayes.html