Aprendizagem Aplicada à Segurança

Mário Antunes

September 22, 2023

University of Aveiro

Table of Contents

SPAM

- The term "spam" is internet slang that refers to unsolicited commercial email (UCE).
- The first reported case of spam occurred in 1898, when the New York Times reported unsolicited messages circulating in association with an old swindle.
- The term "spam" was coined in 1994, based on a now-legendary Monty Python's Flying Circus sketch, where a crowd of Vikings sings progressively louder choruses of "SPAM! SPAM! SPAM!"

THE SPANISH PRISONER

The New York Times reports of unsolicited messages circulating in association with an old swindle.

1898

1 ARPANET

The first reported case of email spam is attributed to Digital Equipment Corporation and circulated to 400 ARPANET users.

POST MAIL

Advertisement based on unsolicited content has been mailed to our doors by Post Mail services for over a century!

2 Early 1900s

3 SEARCH ENGINES

Web content spam and link farms are common forms of spamdexing, the manipulation of Web search result ranking.

1978

THE EMAIL EPIDEMIC

A growing fraction of emails is spam. Platforms and ISPs start investing in spam filtering techniques.

4 Mid 1990s

5 FAKE REVIEWS

Giants of e-commerce like Amazon and Alibaba fight the manipulation of product popularity by opinion spam.

1995

SOCIAL NETWORKS

The rise of Facebook, Twitter, and Reddit leads to new opportunities for spammers to reach billions of Social Web users.

6 2000s

7 SOCIAL BOTS

Millions of accounts operated by software populate social media to carry out nefarious spam campaigns.

Early 2000s

PHISHING

Social engineering and disguise may allow attackers to trick victims into revealing sensitive information. Ransomware are used to extort funds from the victims.

8 Mid 2000s

9 AI SPAM

Systems based on AI can manipulate reality, producing indistinguishable alternatives. AIs can also be target of manipulation and spam to elicit behaviors of the AI system or of its users.

Early 2010s

FALSE NEWS

Spam Websites are created to deliberately propagate false news related to politics, public health, and social issues.

10 Mid 2010s

2018+

Dear Sir,
I am prince ⬛⬛⬛ from Nigeria. Your help would be very appreciated.
I want to transfer all of my fortune outside if Nigeria due to a frozen account,
If you could be so kind and transfer small sum of 3 500 USD to my account,
I would be able to unfreeze my account and transfer my money outside of
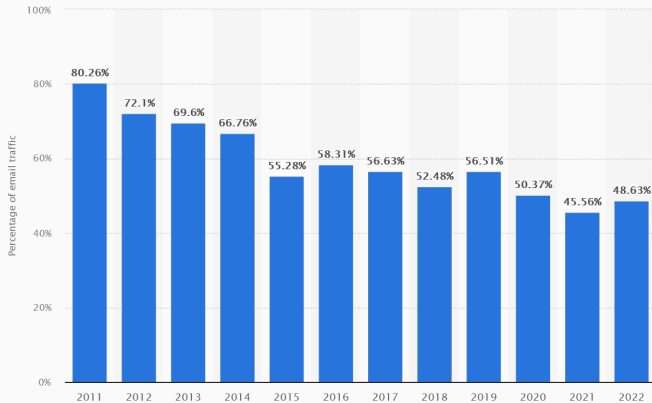Nigeria. To repay your kindness, I will send 1 000 000 USD to your account.
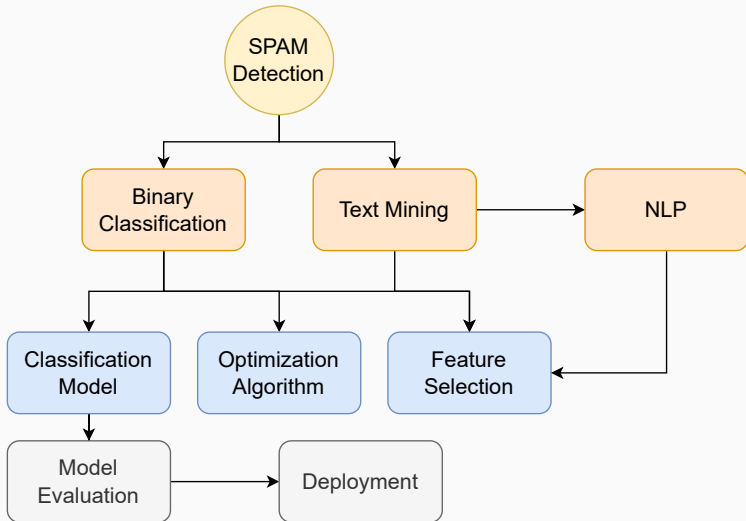
Please contact me to proceed

Prince ⬛⬛⬛

Fight against SPAM

- *Huge* list of `https: //en.wikipedia.org/wiki/Anti-spam_techniques`
- From common sense to *Bayesian spam filtering*
- Unfortunately it is a costly battle

- Binary classification is the task of classifying the elements of a set into two groups (each called class) on the basis of a classification rule.
- For this application one message can either be spam or ham.
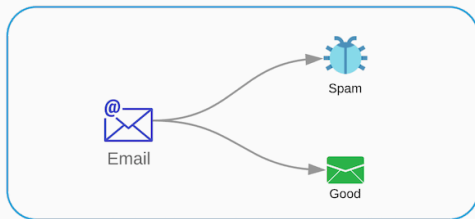
Binary classification

Text Mining

- Text mining is the process of deriving high-quality information from text.
- Combines concepts from Machine Learning, Linguistic and statistical analysis.
- In this area we will explore the methods used to rank words/tokens and the BoW model.

Bag of Words (Bow) model

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Natural Language Processing (NLP)

- NLP gives the computers the ability to understand text.

- Combines *Sintax* and *Semantic* into the analysis.

- One famous exemples are the Large Language Models (LLMs) that power OpenAI Chat GPT.

- SPAM detection is "considered" a toy example.
- As such, we will explore two of the simples learning models: Naive Bayes and Logistic Regression.

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Likelihood of the Evidence given that the Hypothesis is True

Prior Probability of the Hypothesis

Posterior Probability of the Hypothesis given that the Evidence is True

Prior Probability that the evidence is True

$y = b_0 + b_1 x$ ⟸ Linear Model

Logistic Model

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$
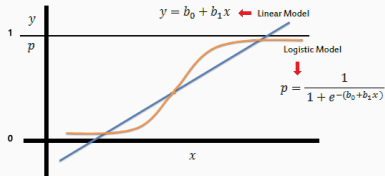
- Classification model can be evaluated using a confusing matrix
- The simplest methods to evaluate a model is throuhgh accuracy: $acc = \frac{TP+TN}{TN+TN+FP+FN}$

	Predicted Positive	Predicted Negative	
Actual Positive	**TP** *True Positive*	**FN** *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	**FP** *False Positive*	**TN** *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Model Evaluation