

# Aprendizagem Aplicada à Segurança

---

Mário Antunes

September 22, 2023

University of Aveiro

# Table of Contents

SPAM

SPAM Detection


Binary Classification

Text Mining

Natural Language Processing (NLP)

Classification Model

Model Evaluation

- 
- 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!”

# SPAM



Dear Sir,

I am prince **George Okeke** 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,

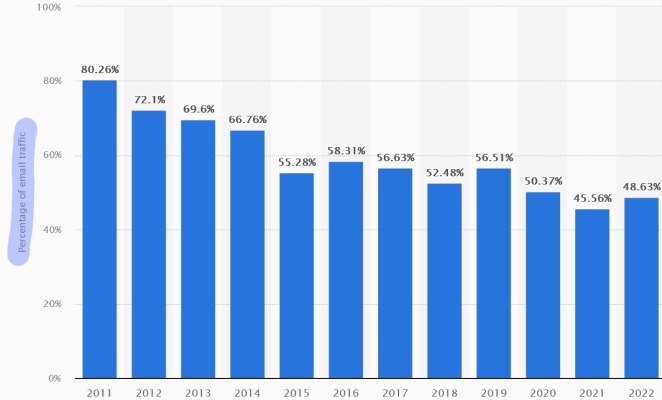
I would be able to unfreeze my account and transfer my money outside of Nigeria. To repay your kindness, I will send 1 000 000 USD to your account.

Please contact me to proceed

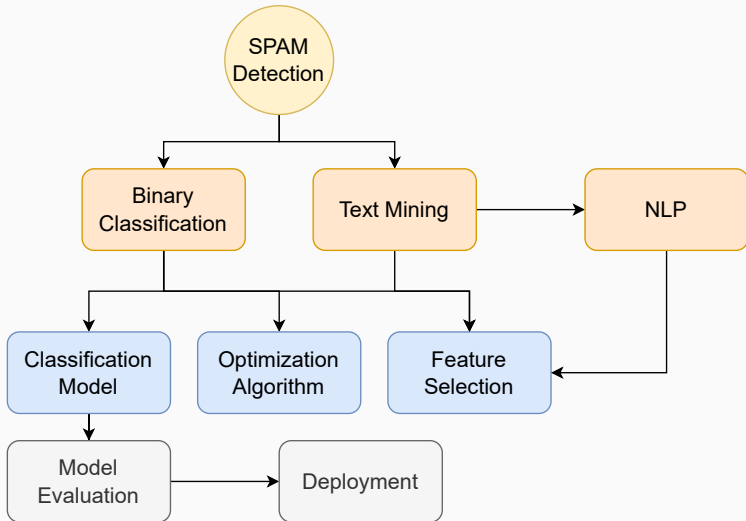
Prince **George Okeke**

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

# Fight against SPAM



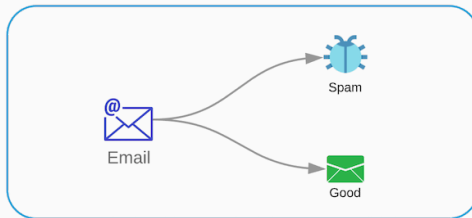
# SPAM Detection





# Binary Classification

- 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 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 *Syntax* and *Semantic* into the analysis.
- One famous examples are the Large Language Models (LLMs) that power OpenAI Chat GPT.

# Classification Model

- SPAM detection is “considered” a toy example.
- As such, we will explore two of the simplest learning models: Naïve Bayes and Logistic Regression.

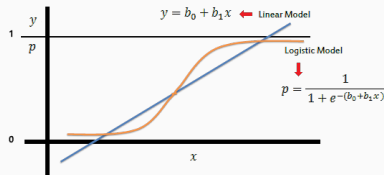
Likelihood of the Evidence given that the Hypothesis is True

Prior Probability of the Hypothesis

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

Posterior Probability of the Hypothesis given that the Evidence is True

Prior Probability that the evidence is True



# Model Evaluation

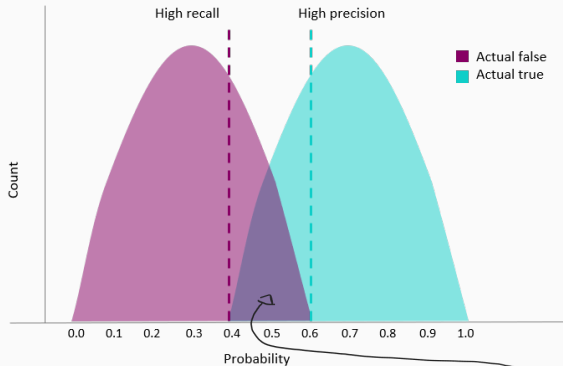
- Classification model can be evaluated using a confusing matrix

- The simplest methods to evaluate a model is through

accuracy:  $acc = \frac{TP+TN}{TP+TN+FP+FN}$

	Predicted Positive	Predicted Negative	
Actual Positive	TP <i>True Positive</i>	FN <i>False Negative</i>	Sensitivity $\frac{TP}{(TP + FN)}$
Actual Negative	FP <i>False Positive</i>	TN <i>True Negative</i>	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



A spam AI can't ever be 100%, because there's sometimes info that is false and true.

