Anti-Spam detection and interpretation

Marcin Łukaszyk, Jean-Baptiste Soubaras

NLP Project 2022/2023

Mr Mark Lukunku <wendyarandas@sheldonisd.com>

I request your approval to present you as next of kin to pending inheritance funds to the sum of USD17.3 Million in my bank. If this suits your interest, please revert back and for more additional information.

I await your reply,

Mr. Mark Lukunku

What is spam ?

Unsolicited e-mail sent to a large number of people with a purpose of:

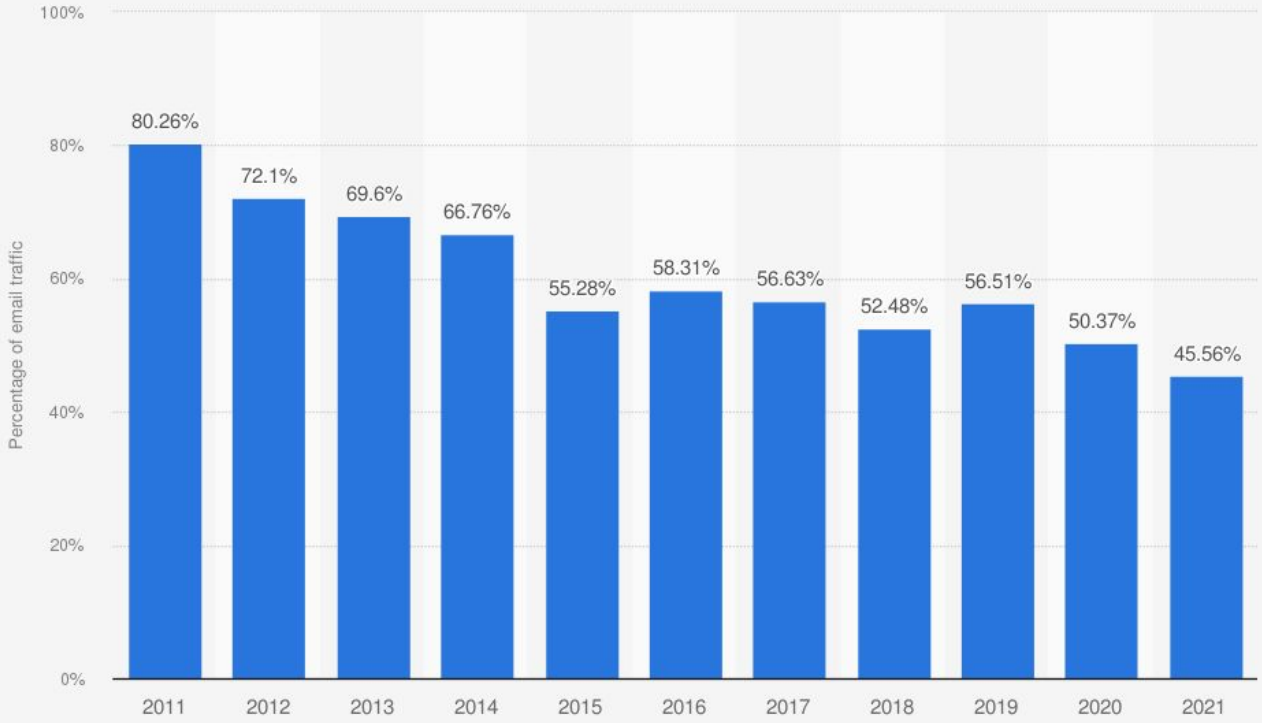
- Advertising
- Phishing
- Spreading malware
- etc…

A big issue

- 80% of the e-mail traffic in 2011

- Costs businesses $20.5 billion every year

Global spam volume as percentage of total e-mail traffic from 2011 to 2021



Source
Kaspersky Lab

Additional Information:
Worldwide; Kaspersky Lab; 2011 to 2021

Performant AI-based filters…

- Development of RNN and NLP

- Google’s filter → 99.9% accuracy

… but opaque

- Which criterions or features ?
- No insights on the global problem
- Hard to detect weaknesses or optimization axis

→ **Need for XAI** (*Explainable Artificial Intelligence*)

Problem Statement

Solve the spam detection problem with XAI

→ Apply recent work on XAI on classical AI methods

Spam filtering algorithms

- Naive Bayesian algorithm
- K-Nearest Neighbors
- Reverse dbscan
- Multi-Layer Perceptron
- Convolutional Neural Network
- LSTM

Spam filtering algorithms

Naive Bayesian algorithms

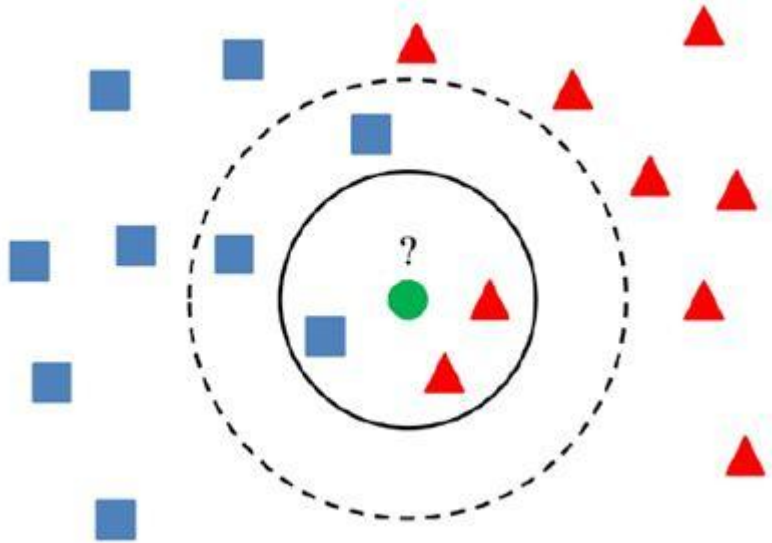
Counting probability of email being spam based on frequency of cointened words being in spam messages

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

Spam filtering algorithms

K-Nearest Neighbors

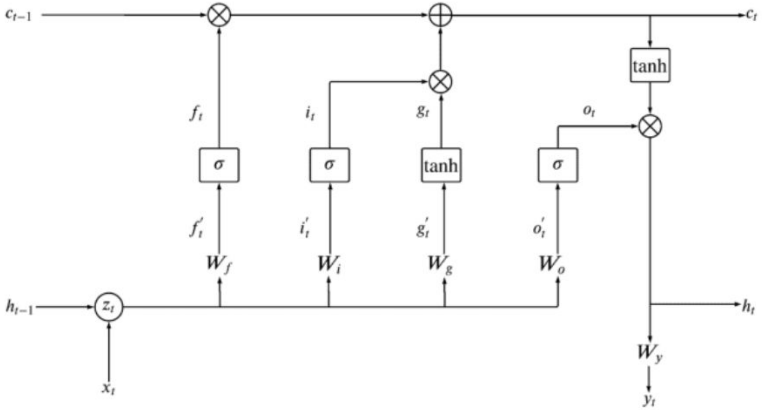
Classification based on value of k closest values in dataset.
To determinate distance we can count similar words or use cosine similarity.



Spam filtering algorithms

Deep Learning Methods

Multi-Layer Perceptron, Convolutional Neural Network or Long Short Term Memory networks



XAI methods

- Linguistic Correlation Analysis
- Cross-model Correlation Analysis
- Auxiliary prediction

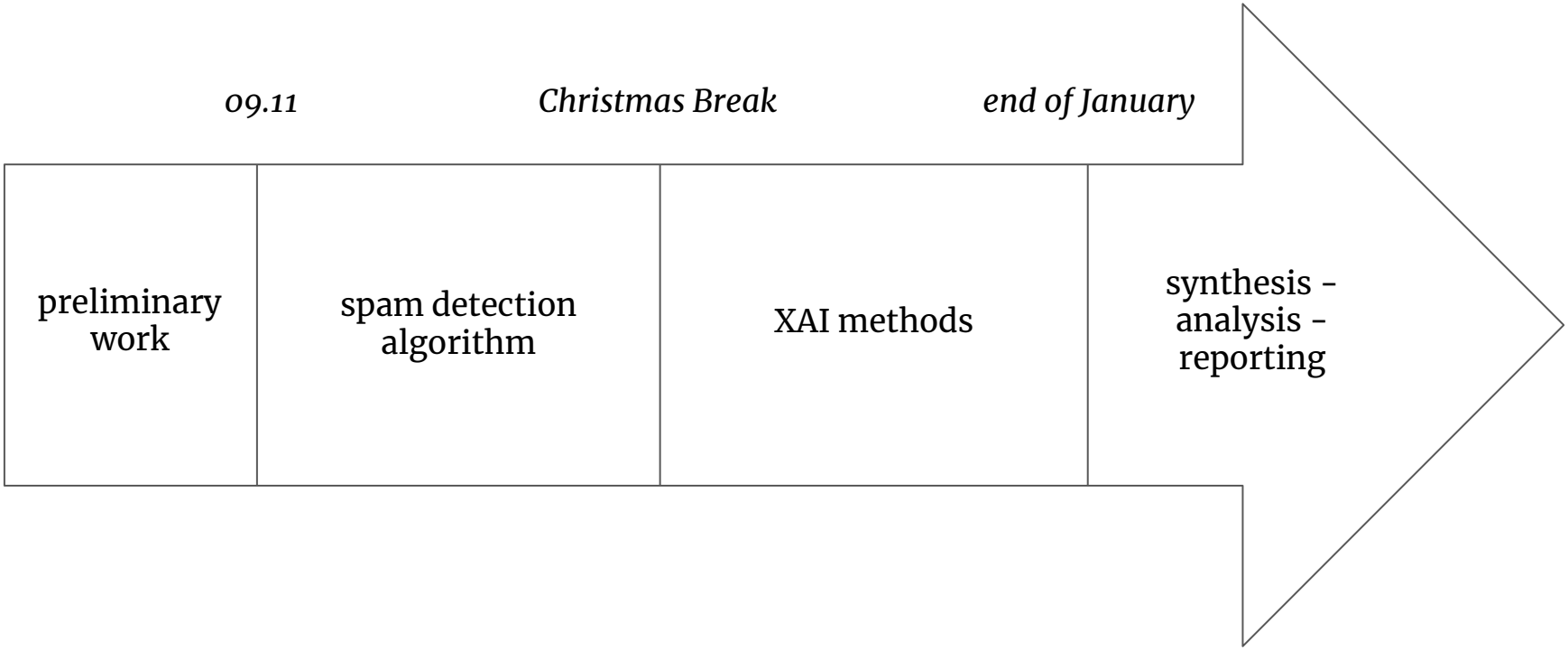
Methodology

- Work environment implementation
- Writing a AI-based spam classifier
- Combining at least two different XAI methods with the algorithm
- Analysis and synthesis

Questions

Work Plan

1. Bibliographical Research / Preliminary Work
2. Implementation of a spam detection system based on AI
3. Research and implement XAI methods
4. Synthesize all the insights and deduce a methodology for RNN interpretation


09.11
Christmas Break
end of January
preliminary work
spam detection algorithm
XAI methods
synthesis - analysis - reporting