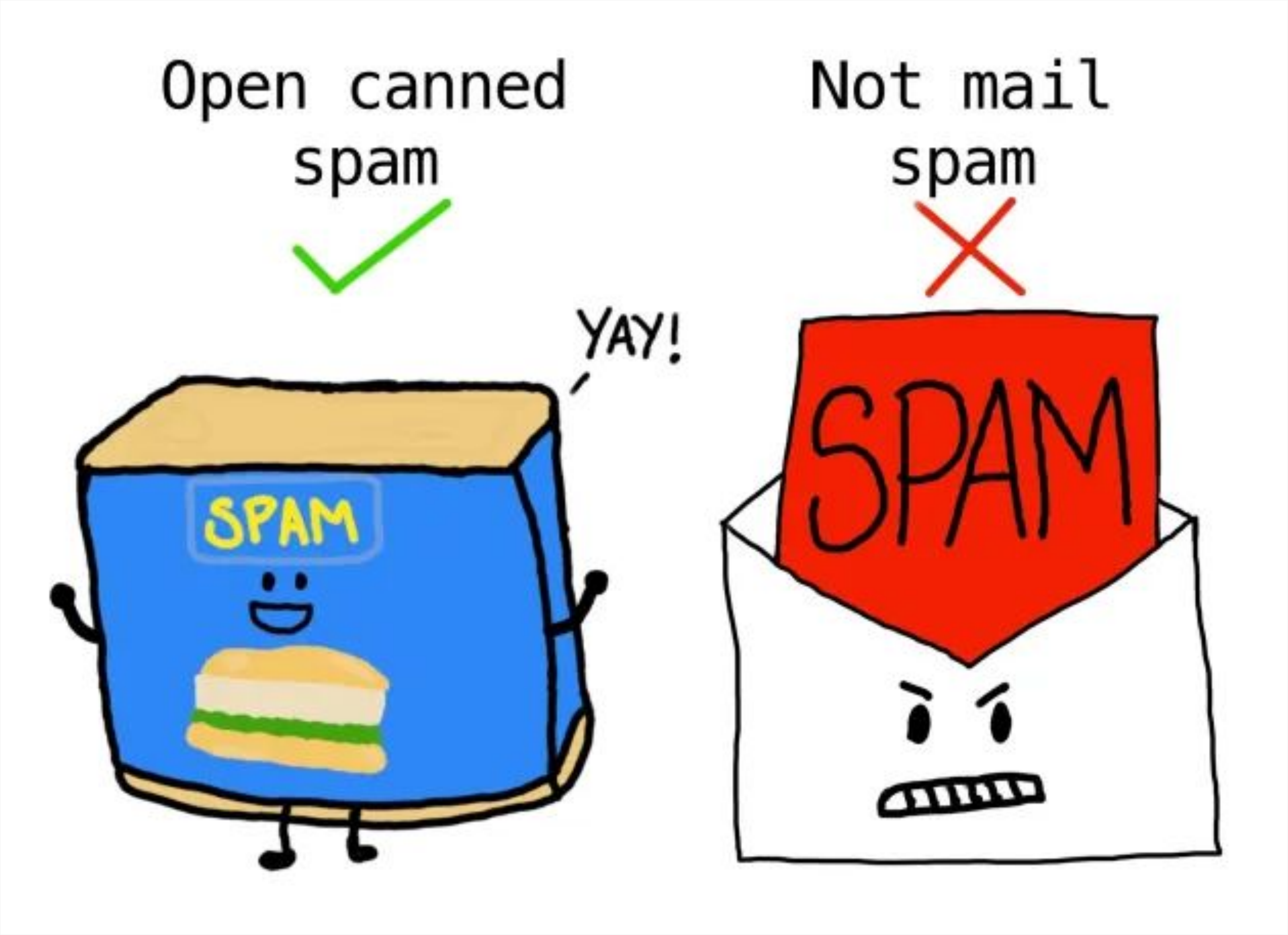
Spam Detection

Kevin Zhuo & Daniel Zhang



Introduction

Spam or Real?

- “Night has ended for another day, morning has come in a special way. May you smile like the sunny rays and leaves your worries at the blue blue bay. Gud mrng”

- “URGENT This is our 2nd attempt to contact U. Your å£900 prize from YESTERDAY is still awaiting collection. To claim CALL NOW 09061702893”

- “U GOIN OUT 2NITE?”

Deciding between spam and non-spam SMS messages is a skill that our generation has honed over the years. However, for other generations, detecting spam is not as straightforward. A computational model can help people of all ages protect against cybersecurity attacks, avoid phishing attacks, and block irrelevant notifications.

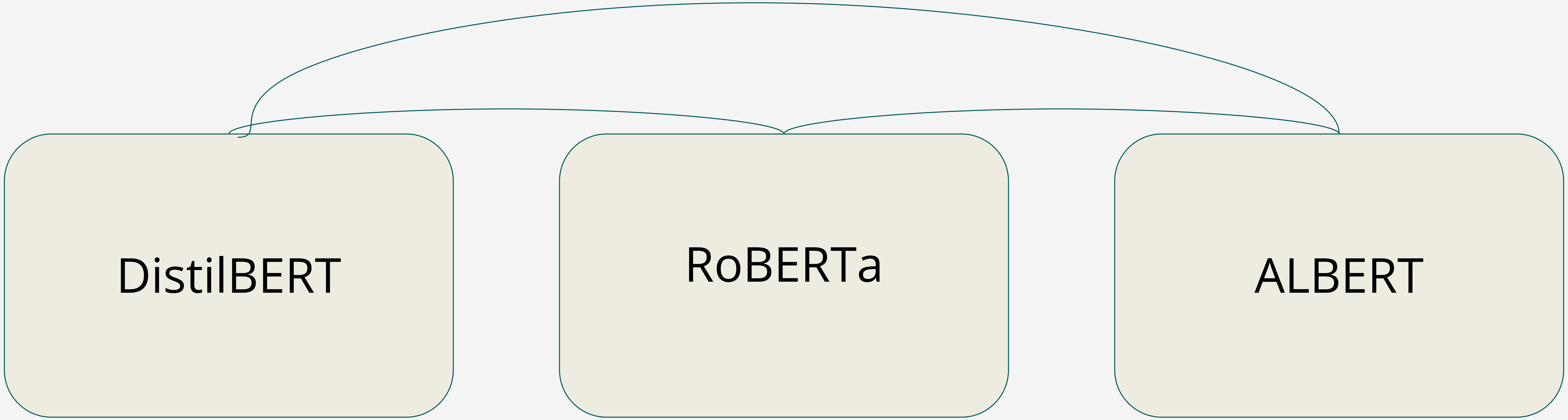
Dataset

The previous examples were real examples from our dataset: SMS Spam Collection.

- 5,574 messages, tagged either ham or spam.
  - Imbalanced Data (747 spam messages and 4827 ham messages)!

Graphs depicting token distribution for initial and training dataset.

Proposed Methods



DistilBERT

RoBERTa

ALBERT

Results/Discussion

DistilBERT. It yielded a test accuracy of 0.993028, a f1 score of 0.970711, and a balanced accuracy of 0.971545. Below is a confusion matrix describing DistilBERT's performance.

RoBERTa. It yielded a test accuracy of 0.0.996016, a f1 score of 0.984, and a balanced accuracy of 0.997730. Below is a confusion matrix describing DistilBERT's performance.

ALBERT. It yielded a test accuracy of 0.875498, a f1 score of 0.031007, and a balanced accuracy of 0.505860. Below is a confusion matrix describing DistilBERT's performance.