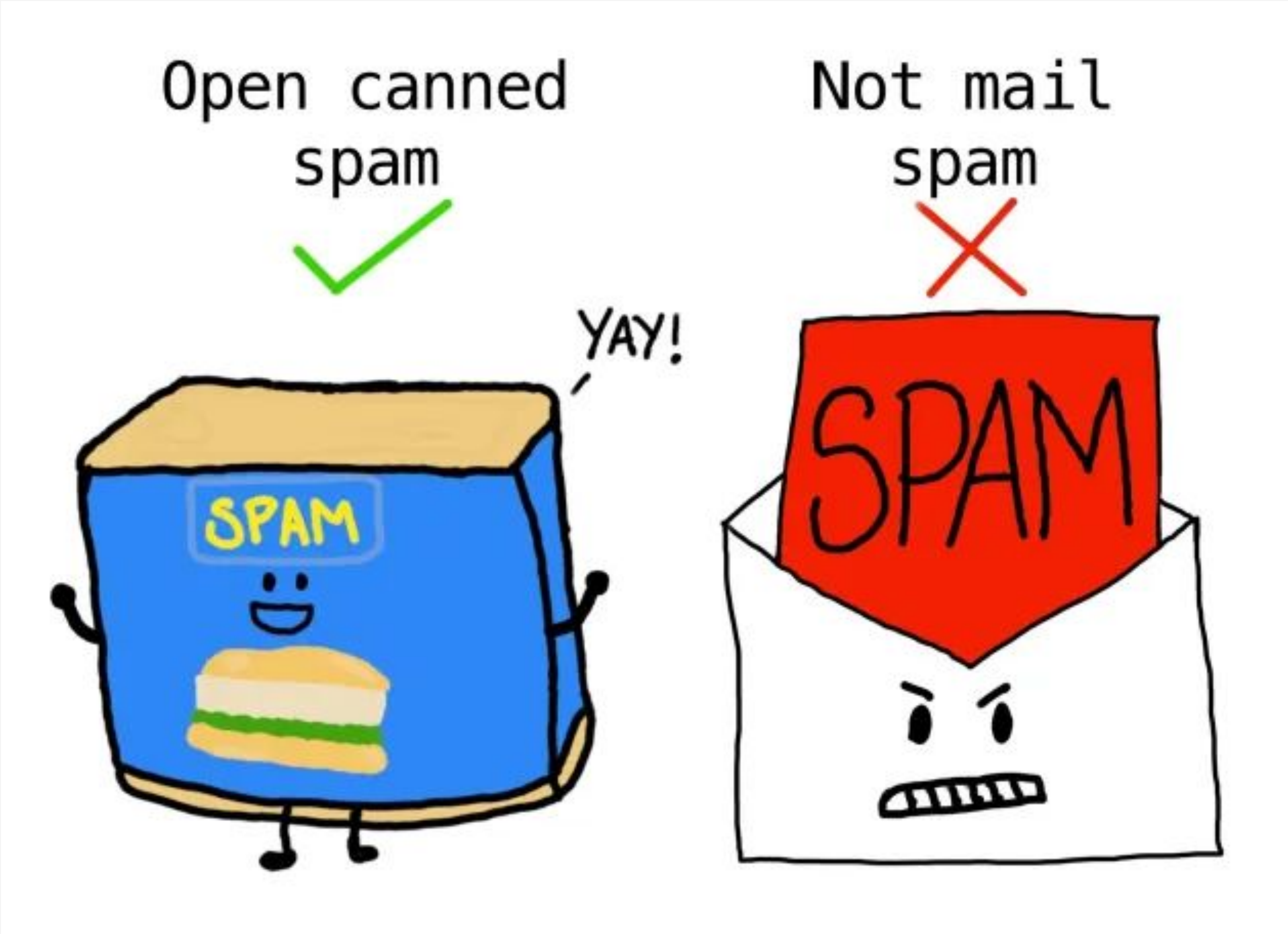
Spam Detection

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.

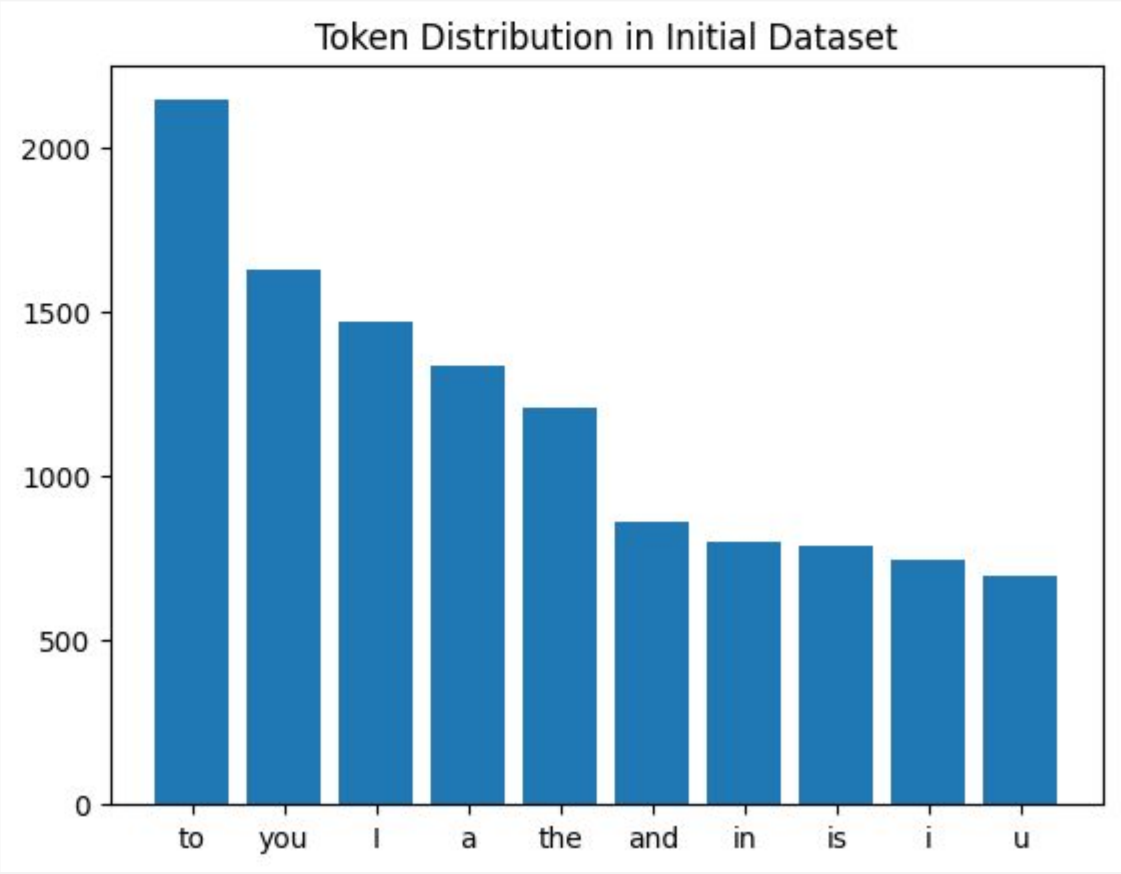
Dataset

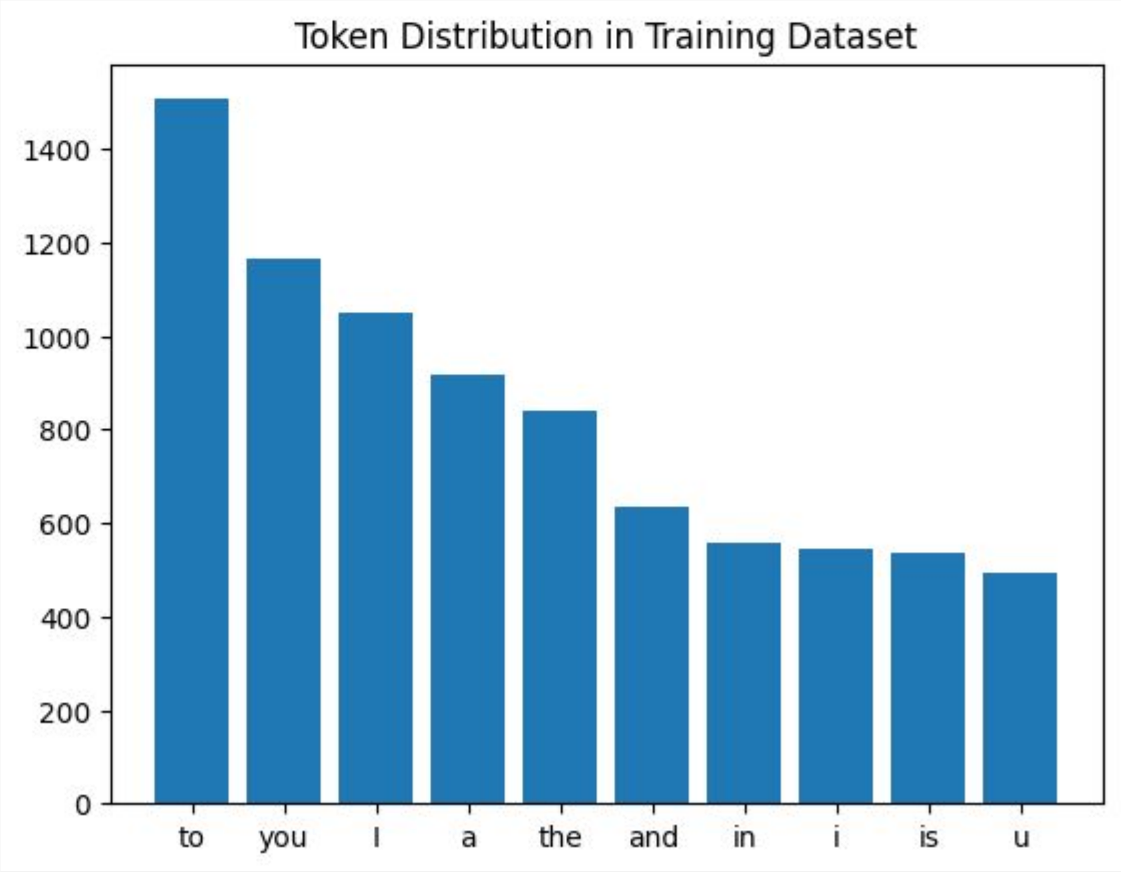
The previous examples were real examples from our dataset: SMS Spam Collection.

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

- Collected from a wide variety of sms messages across numerous sources

- Token Distribution exhibits an exponential distribution





Proposed Methods

DistilBERT

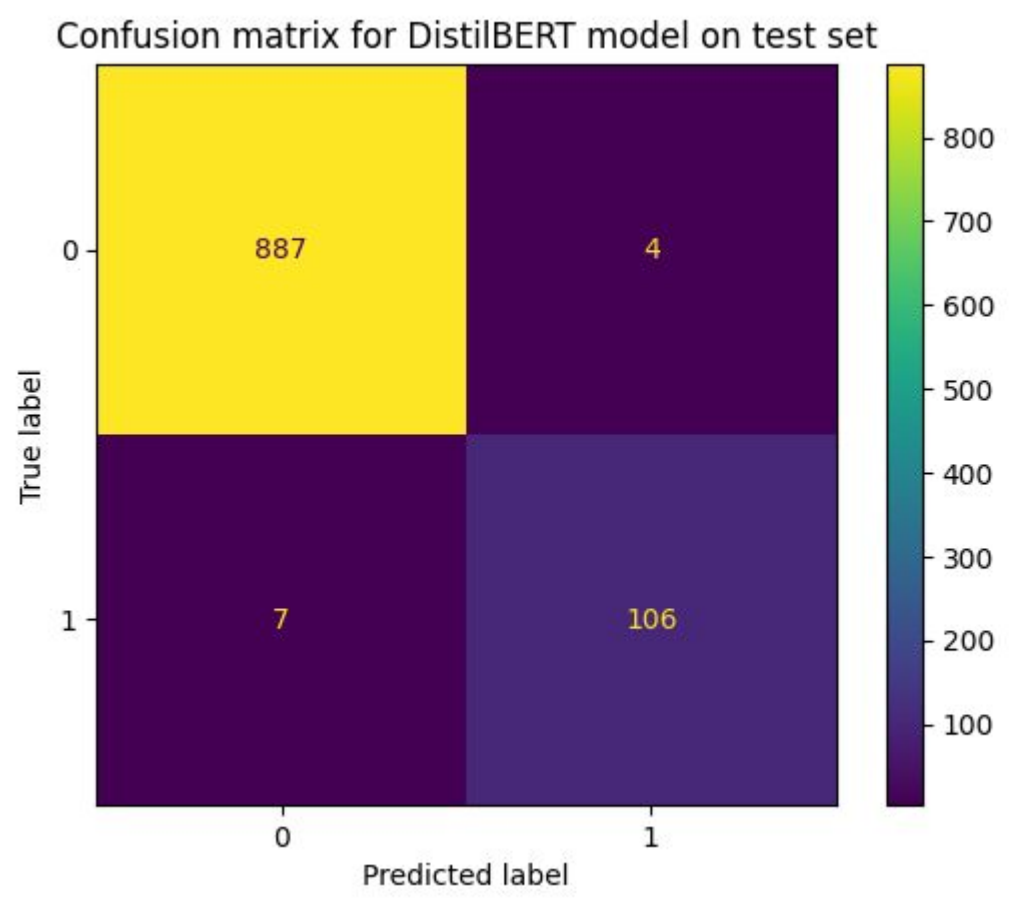
RoBERTa

ALBERT

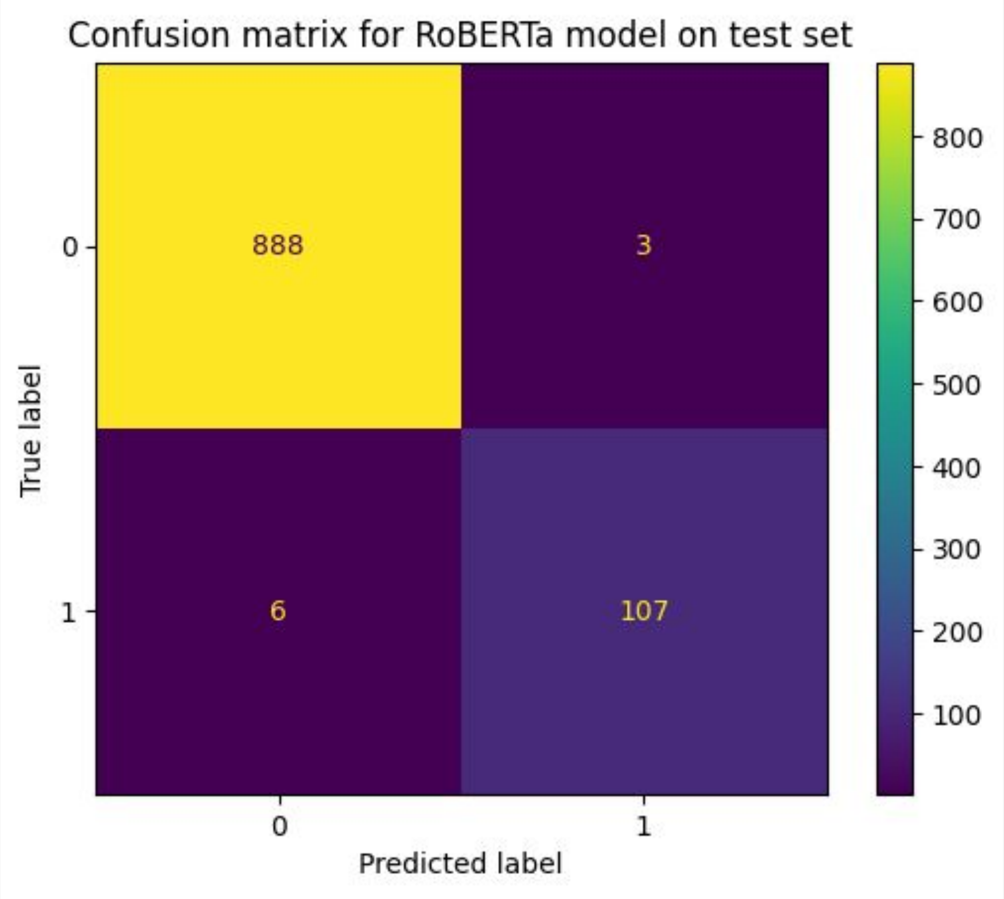
These three models are all inspired by the BERT model architecture. We will train all 3 models on the same training dataset and test all 3 models on the same test set.

Will the results all be the same then?

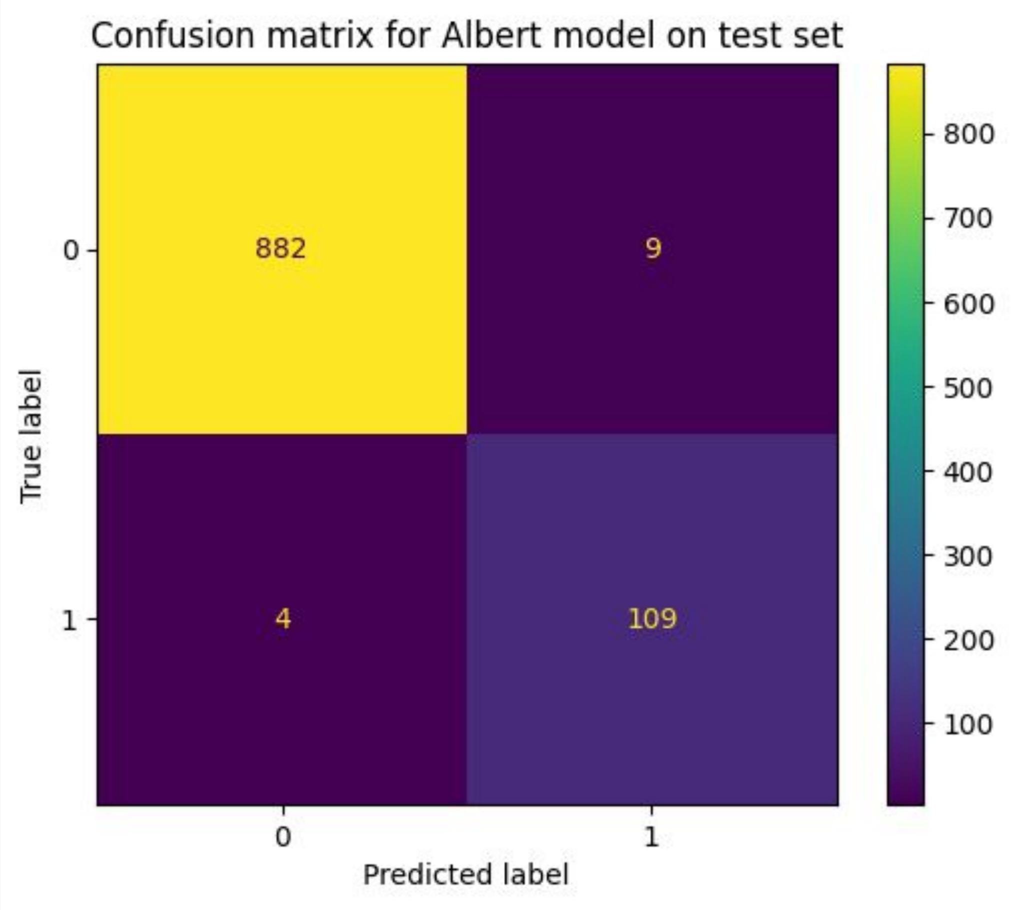
Results/Discussion



DistilBERT
- Test accuracy = 0.98094
- F1 score = 0.95067
- Balanced accuracy = 0.96678



RoBERTa
- Test accuracy = 0.99013
- F1 score = 0.95964
- balanced accuracy = 0.97176



ALBERT
- Test accuracy = 0.98705
- F1 score = 0.94372
- Balanced accuracy = 0.97725