YZV 311E “Data Mining” Term Project

E-Mail Spam Classification

Group 8

- Abdullah Bilici, 150200330
- Bora Boyacıoğlu, 150200310

Growing Importance & Key Limitations

Importance

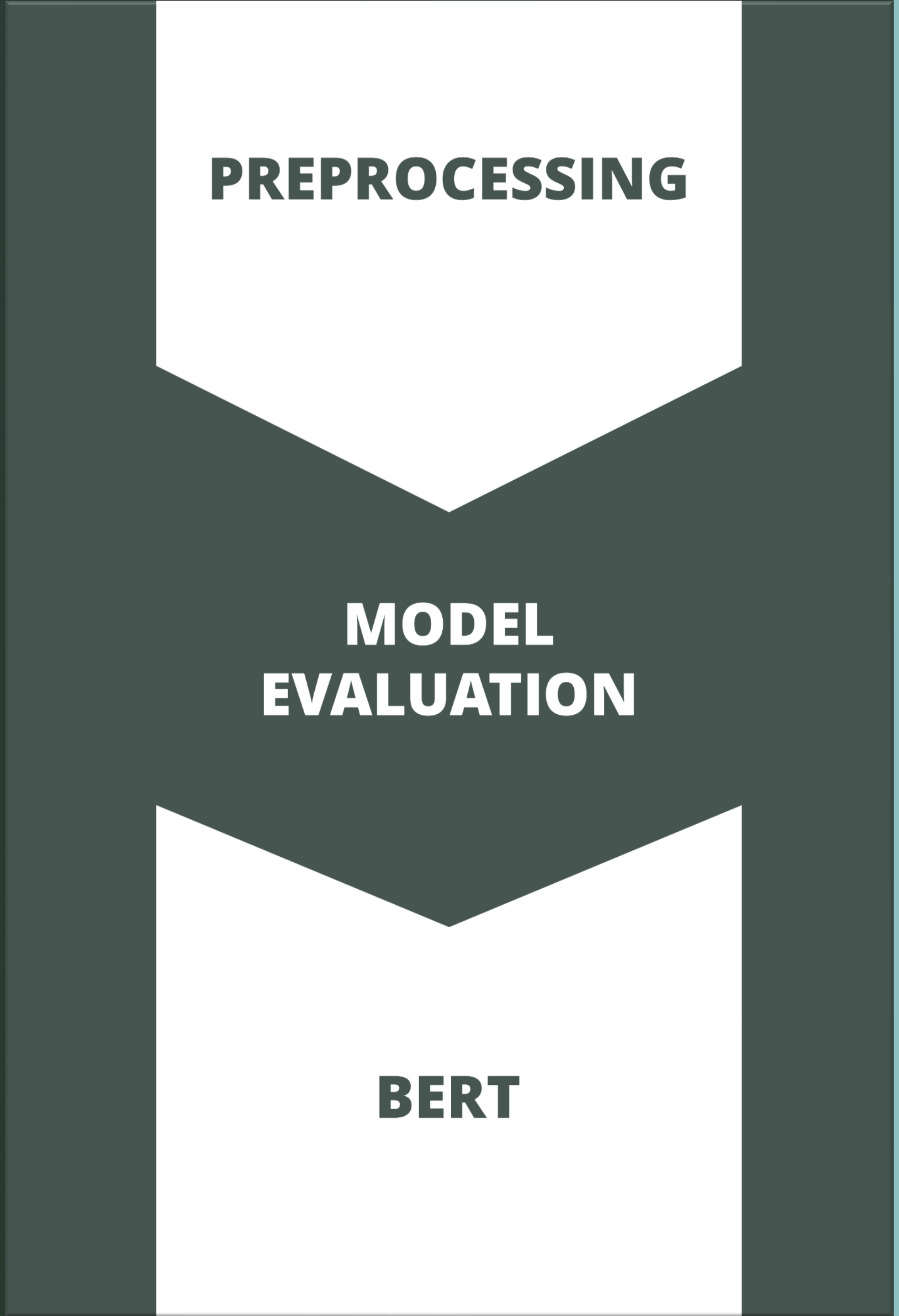
- Scams: Individuals
- Malware: Compaines
- Lots of Junk

Limitations

- Hard to distinguish
- Ineffective Algorithms
- False Categorising

Highlights

- Filtering the punctuations, stop-words, etc. after vectorising
- Using different models and comparing them for the model
- Hyperparameter tuning
- Experimenting on a special **BERT** model



PREPROCESSING

MODEL EVALUATION

BERT

Related Studies

There are important researches and competitive works around for this task.

- Big companies like Google, Apple and Microsoft for personal addresses
- Smaller companies for their business related services

- Kaggle competition in 2011
- Our dataset’s model: Harsh Sinha

System Architecture:
Pre-Processing

Tokenising

- Removing stop words and punctuations
- Lowercasing
- Lemmatisation

Feature Modelling

- TF-IDF
  - Term Frequency
  - Inverse Document Frequency
- Feature Reduction

System Architecture:
Model Creation

Models

- Multinomial Naive Bayes
  - A: 0.86, P: 1.00, R: 0.40, F1: 0.58
- Support Vector Machines
  - A: 0.95, P: 0.99, R: 0.80, F1: 0.89
- Random Forest
  - A: 0.98, P: 0.98, R: 0.93, F1: 0.96
- Logistic Regression
  - A: 0.90, P: 1.00, R: 0.58, F1: 0.74

Hyperparameter Tuning

- Random Forest
  - `n_estimators: 400`
  - `min_samples_split: 2`
  - `min_samples_leaf: 1`
  - `max_features: Square Root`

System Architecture:

BERT

Model Creation

- Creation
- Training
- Loop

Evaluation

- Accuracy: 0.98
- Precision: 0.96
- Recall: 0.94
- F1 Score: 0.95

Dataset

- Our dataset consists of two columns: `text` and `spam`, which contains the email content and an indicator for spam/ham.
- There are 5695 rows.
- %76 of them are ham, the rest are categorised as spam.

Experimental Results

Random Forest

- After hyperparameter tuning, the results are:
  - `Accuracy : 0.97`
  - `Precision: 0.97`
  - `Recall   : 0.94`
  - `F1 Score : 0.95`
- Using the following properties:
  - `n_estimators: 400`
  - `min_samples_split: 2`
  - `min_samples_leaf: 1`
  - `max_features: Square Root`

BERT

- With 20 epochs and 0.0002 learning rate, the results are:
  - `Accuracy : 0.98`
  - `Precision: 0.96`
  - `Recall   : 0.94`
  - `F1 Score : 0.95`

Conclusions
& Thanks

- We got pretty high results, which seems to be the maximum with our limitations.
- There are of course more ways to improve these, with much complex models.
- Thank you for listening to our presentation.