Reinforcement Learning for Email Security Problem

The target of the Reinforcement Learning model is to successfully detect whether an email is safe or if it is a phishing email.

Dataset

The dataset used is available at the following Kaggle link. This dataset has the following properties:

Columns: [Email Text, Email Type] Rows: 18650

Thus, this dataset contained 18650 examples of both phishing and safe emails, containing the full email text for each of the emails.

Model

As the full email text had been provided in the dataset, it was concluded that a transformer model should be used to generate embeddings for the emails, which can then be fit into a regular Reinforcement Learning model.

Thus, the models used for this project are:

1. BERT for Natural Language Processing
2. Deep Q Learning for Reinforcement Learning

Methodology

The methodology of this project is as follows:

3. Import all required libraries
4. Initialize the BERT transformer model
5. Load the dataset of emails from kaggle into a Pandas DataFrame
6. Create a new database containing embeddings for each Email Text and categorical classification of the Email Types
7. Define the Reinforcement Learning Environment based on the gym library
8. Define a Deep Q Learning Reinforcement Learning model
9. Train the model on the majority(75%) of the total data
10. Test the model on the remaining data