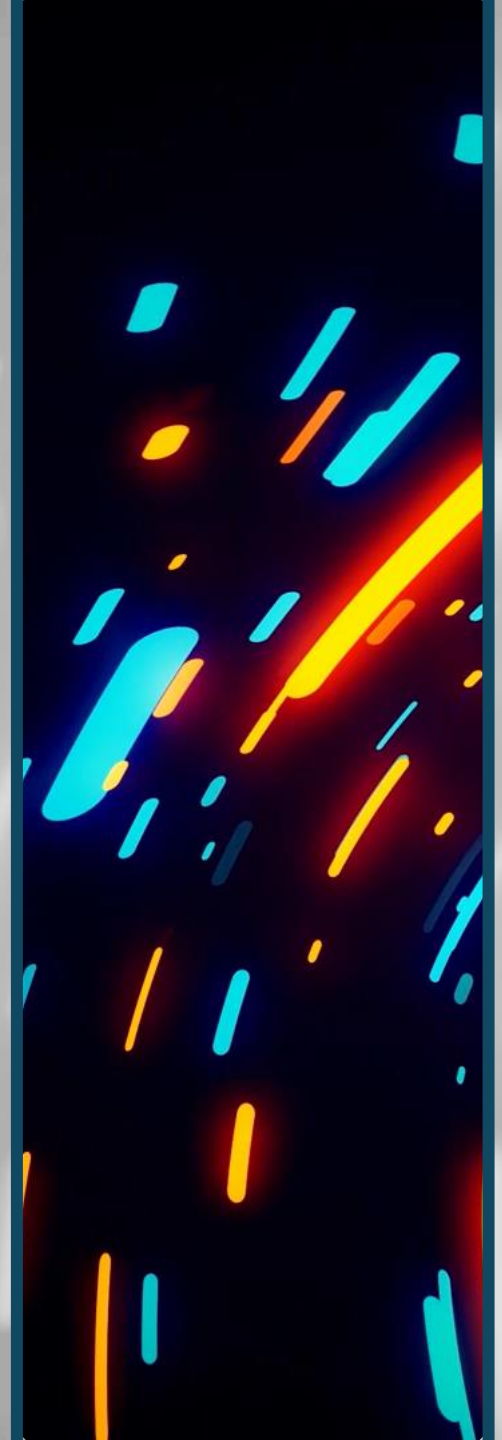


BERT-Based Email Classifier

Date: 20/08/2024

Gal Granot

Nir Tevet

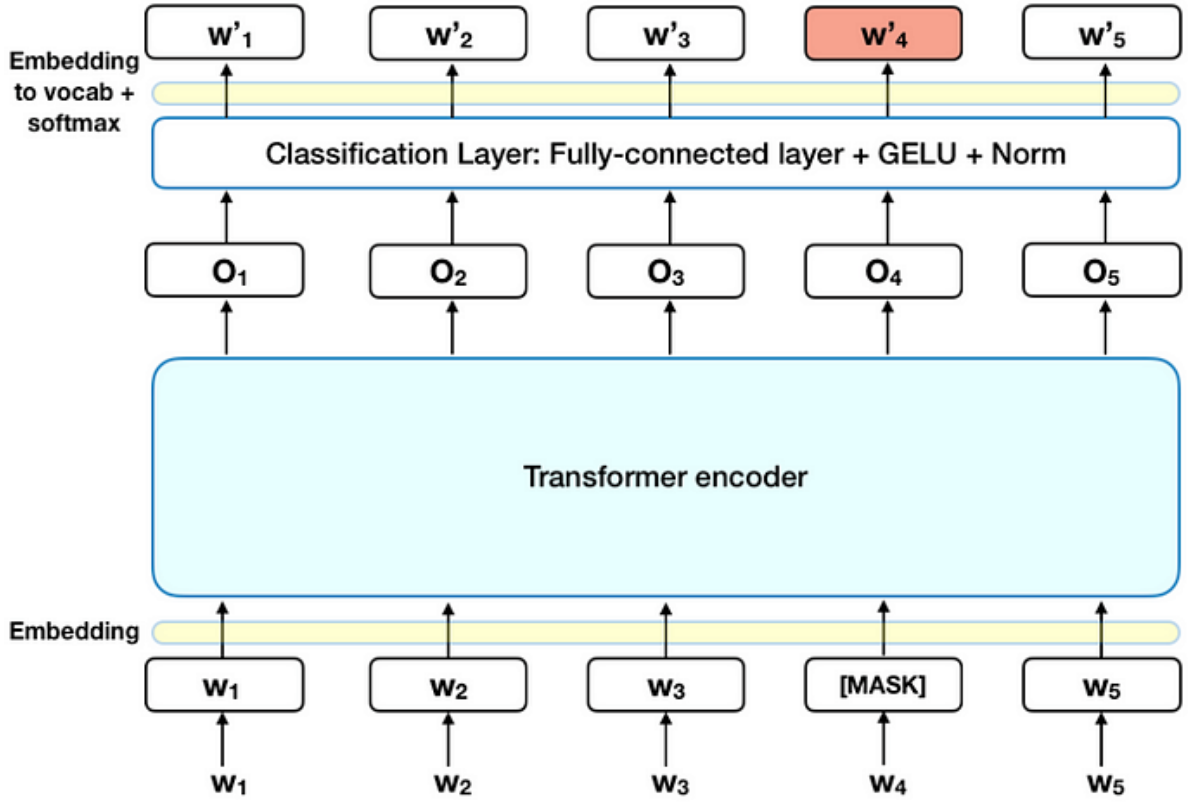
ECE Deep Learning 046211

Overview

- Objective – classifying legitimate emails from phishing attempts while retaining explainability to the user
- Traditional methods (SVM, Random Forests, Naïve Bayes) generally lack deep semantic understanding
- Motivation – phishing is a serious cybersecurity threat, targeting individuals and organizations
- NLP gravitates towards transformers in recent years
- Method – fine-tuning a BERT-based transformer

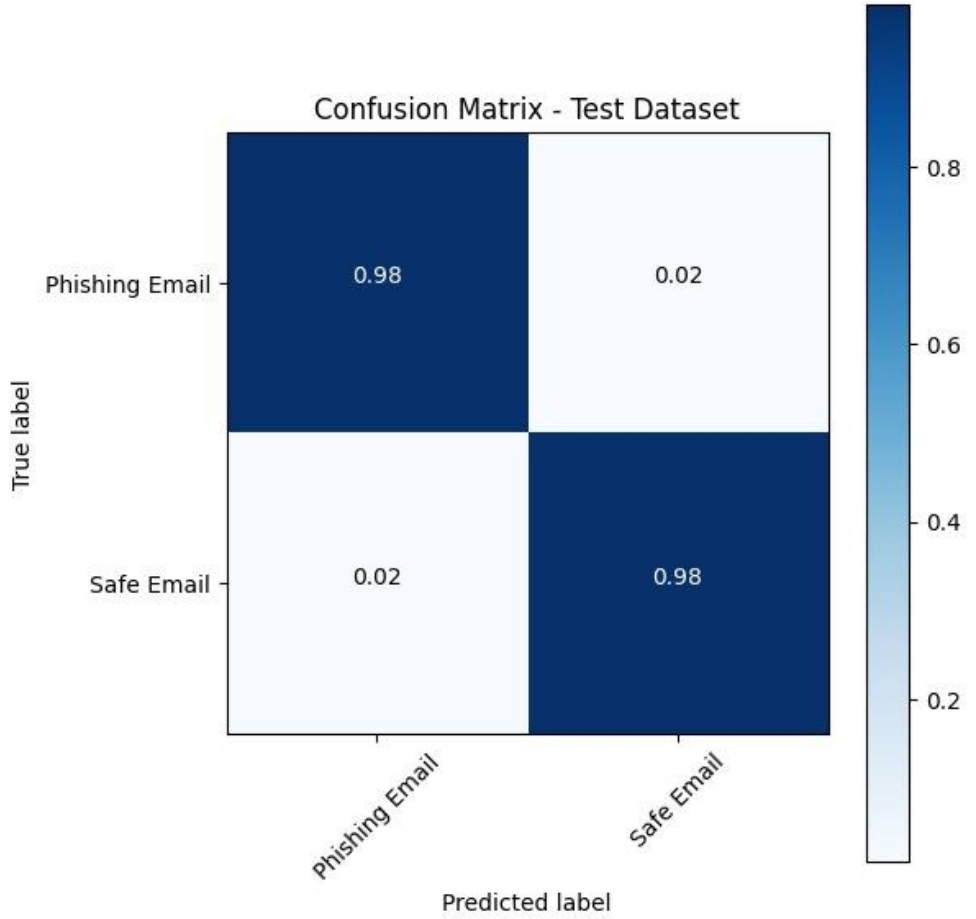
Algorithm Overview

- BERT – Bidirectional Encoder Representations from Transformers
- Based on the attention architecture
- Uses masked language modeling
- Methodology:
  - Preprocessing dataset
  - Model fine-tuning
  - Hyperparameter search
  - Model training
  - Attention mechanism analysis



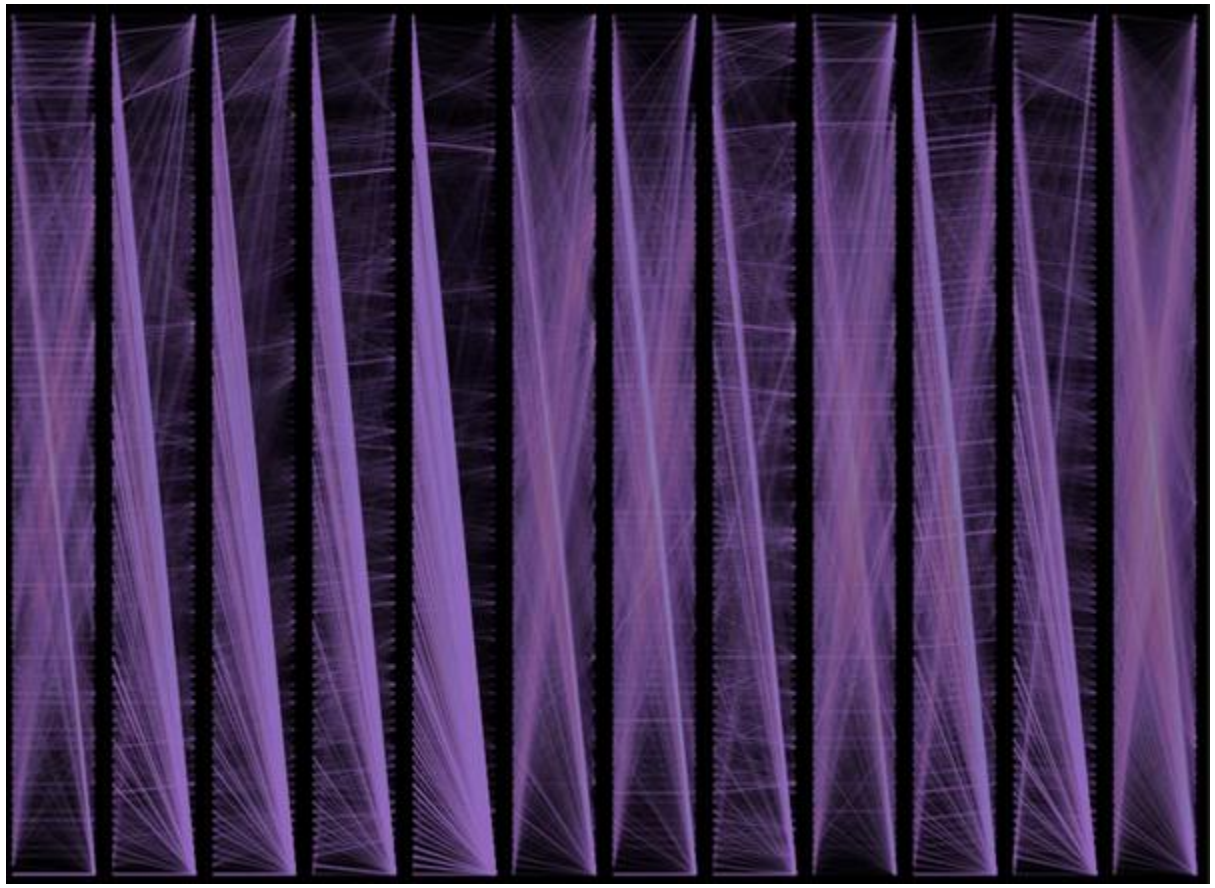
Experimental Results

- 98% accuracy on test dataset
- Minimal and similar false-negatives and false-positives
- Alternative models with classical methods (SVM with TF-IDF) achieved similar test results without the transformer's explainability



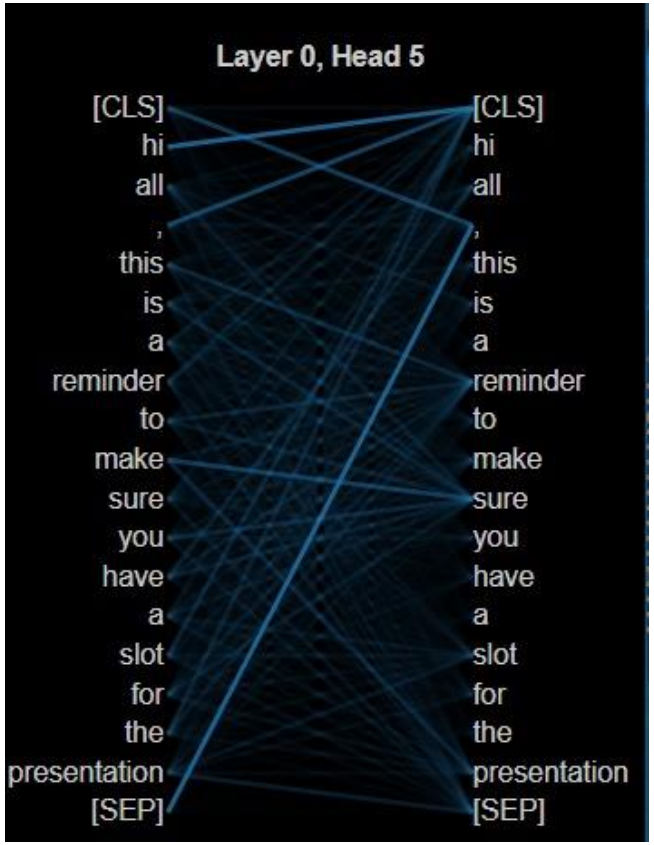
Attention Mechanism

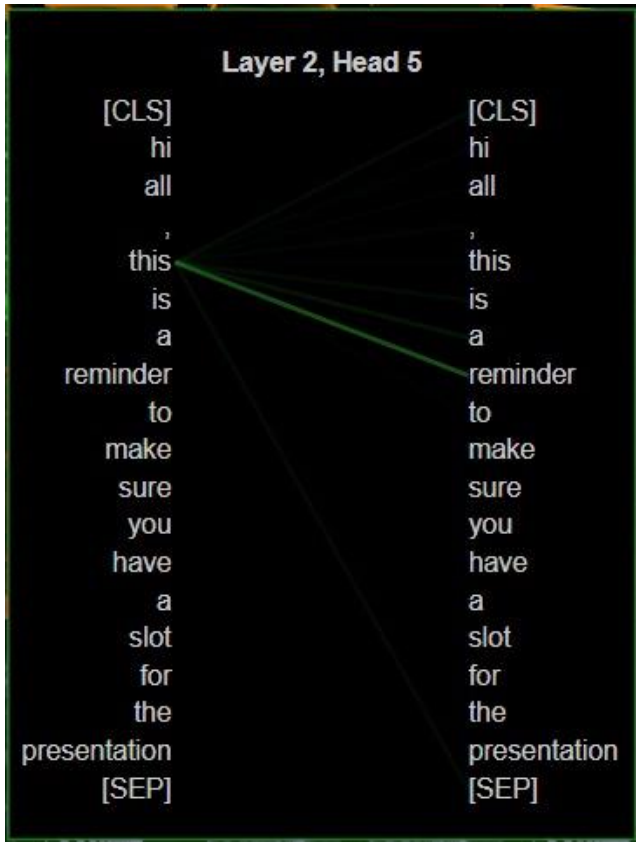
- BertViz – tool for visualizing the attention mechanism, providing explainability to the user
- Notable tokens:
  - [CLS] – prepended to all input sequences by tokenizer, captures overall importance
  - [SEP] – segment divider inserted by tokenizer
  - ".", "()", "," – punctuation used to capture semantic connections between different places in text



Attention Mechanism

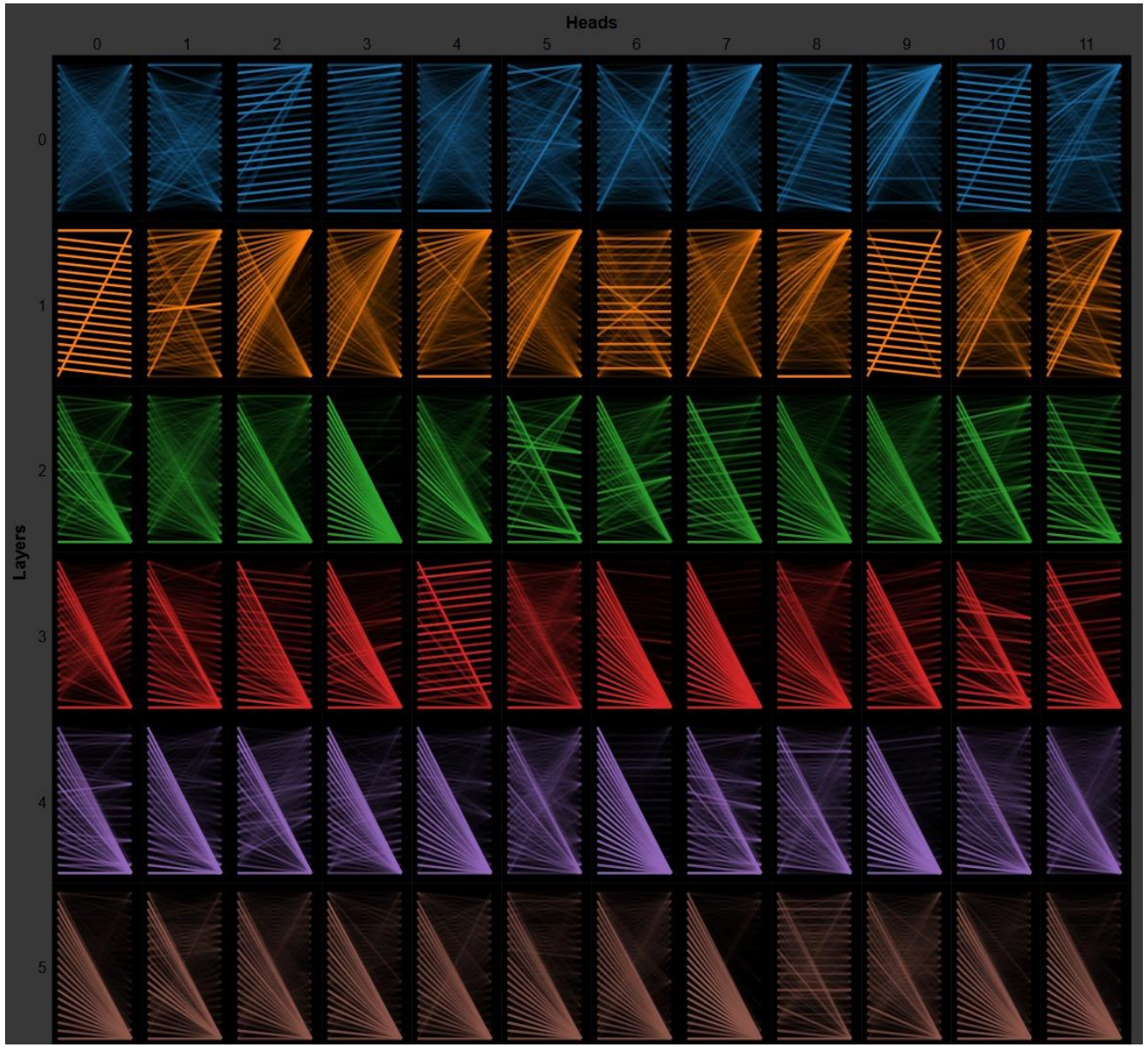
- Each layer learns different semantic relations – close/short term attention
- Important tokens are strongly connected to others
- Some layers learn similar semantic connections and are possibly redundant





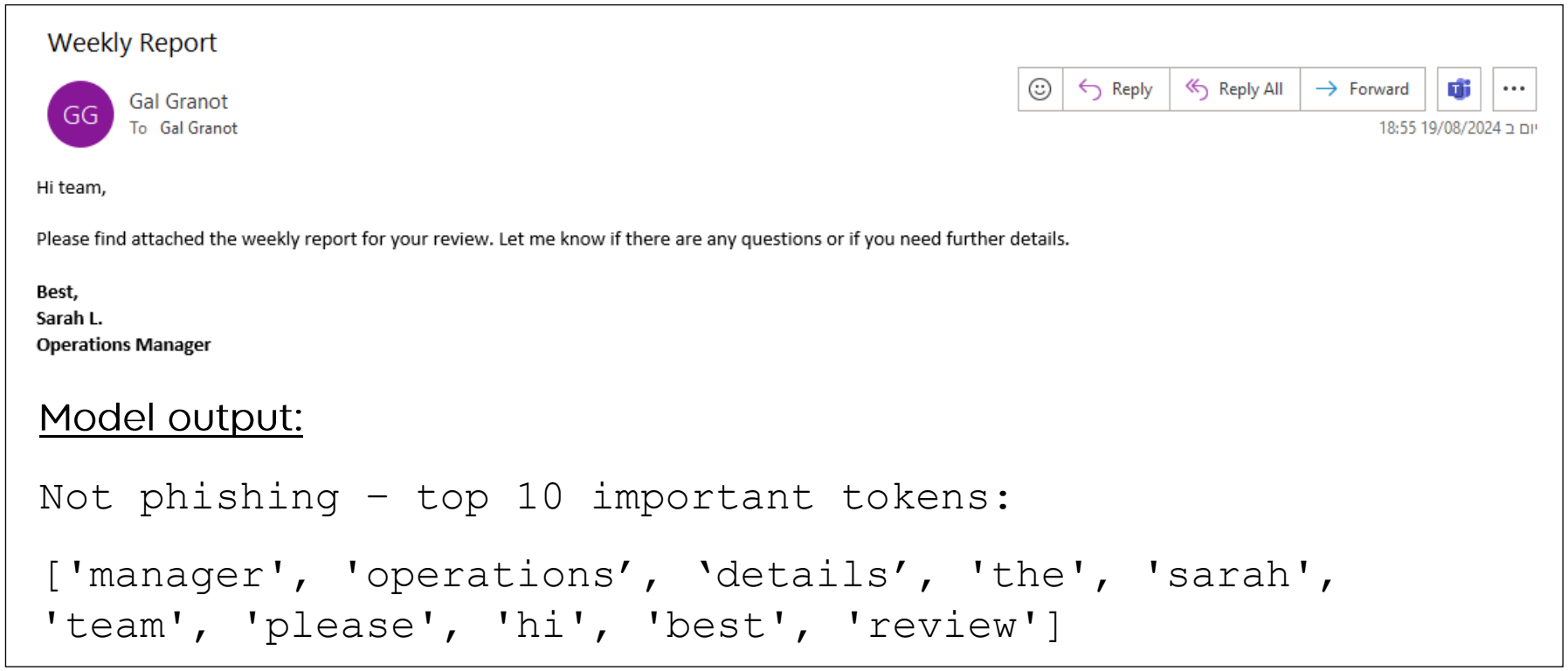
Attention Mechanism

- Each layer learns different semantic relations – close/short term attention
- Important tokens are strongly connected to others
- Some layers learn similar semantic connections and are possibly redundant



Explainability

- By conducting attention analysis provided by model, we can provide explainability to the user
- This is crucial for transparency and user trust



Conclusions

- Fine-tuned pre-trained models' versatility
- Attention mechanism explainability
- Future work:
  - Different setups (optimizers, training regimes, loss functions...)
  - Including classical methods (SVM TF-IDF)
  - Generative email tasks
  - End-to-end privacy implementation

Thank you!