Understanding the semantics of phishing e-mails

Cross-corpora analyses with Large Language Models

Simone Mattia - s.mattia2@campus.unimib.it

Agenda

- Domain
- Project Goals
- Data and Preprocessing
- Large Language Models
- Bias Analyses

Domain

- **Social engineering**, in the context of information security, is the psychological manipulation of people into performing actions or divulging confidential information
- **Phishing** is a specific type of social engineering attack with the aim of acquiring sensitive data, in which the perpetrator masquerades as a legitimate business or reputable person
- Phishing attacks use **e-mail** as their main **vector** and are a **major threat** to information security: the Egress Email Security Risk Report estimates that **92%** of companies have **already** fallen victim to them
- **Understanding** phishing e-mails is therefore **essential** for developing effective **detection and prevention mechanisms**

Egrees - Email Security Risk Report

Project Goals

Propose an approach to detect and quantify the semantic differences between common and phishing e-mails through a cross-corpora analyses based on Large Language Models

Fine tuning of two different models, both based on BERT, using different e-mail corpora and understanding the bias introduced during training using prompting probing techniques

Data and Preprocessing

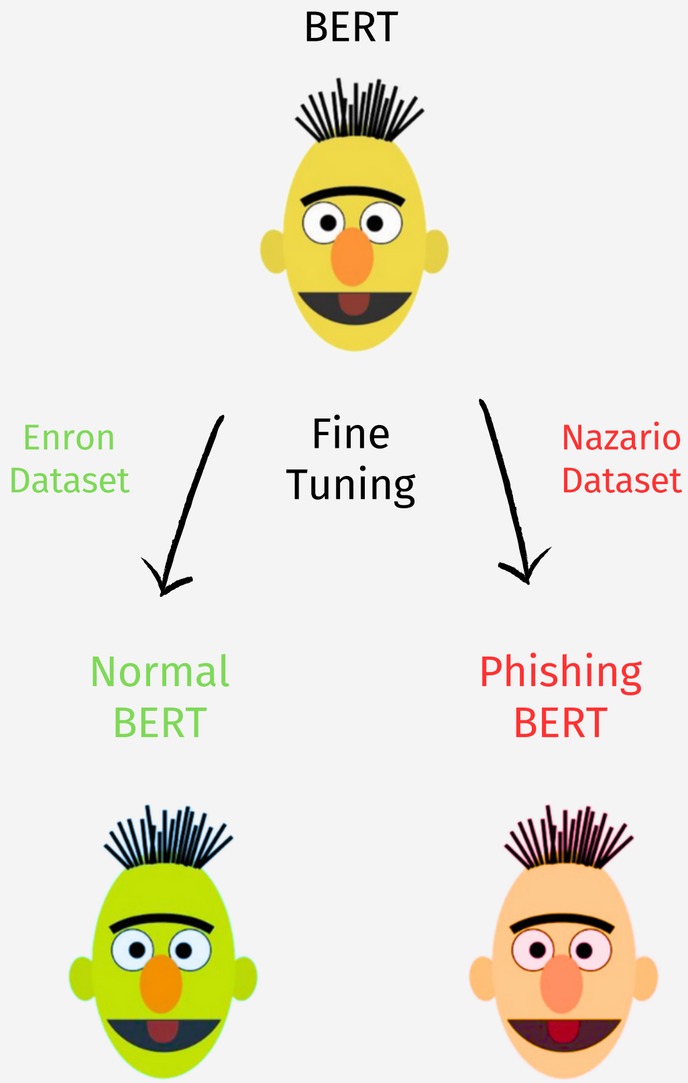
- Data Sources**:**
  - **Enron** Corporate Mails Dataset (500K mails)
  - **Nazario**'s Phishing Mails Dataset (5K mails)

- Preprocessing:
  - Sampling the dataset keeping only 5K mails per type (phishing dataset limit)
  - Replaced urls and e-mail addresses with constant strings
  - Tokenization
    - Minimum number of tokens per sentence: 10
    - Maximum number of tokens per sentence: 512 (BERT limit)

Enron Corporate Mails Dataset
Nazario's Phishing Mails Dataset

Large Language Models

- **Fine-tuning** of two BERT-base model on differente e-mail corpora, each consisting of 69K sentences with 15% masking
- **Training setup**:
  - Batch size: 8 (ColabPro limit)
  - Epochs: 4
  - Learning rate: 5e-5
  - Optimizer: ADAM
- **Training time**: 4 hours, for each model, on ColabPro using an NVIDIA V100



Jacob Devling, 2018, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

Large Language Models

The prediction of some masked tokens shows that there are differences between the two models, especially when considering the domain of phishing mails

	You have to ... me money	Click on links provided in order to ... your ...
	You have to send me money	Click on links provided in order to update your account
	You have to save me money	Click on links provided in order to view your feedback

Bias Analyses

- **Probing** is an approach to understand what linguistic information is contained in the representations of pre-trained language models, it can also be formulated as a prompting task: in that case we refer to **probing via prompting**
- We have several approaches to quantifying some social biases, e.g. gender pronoun resolution for gender bias, but in the context of phishing e-mails it is complex to define a rigorous method
- The literature shows that it is possible to define models that classify the content of an email according to the most relevant personality trait that emerges from the text

Keita Kurita, 2019, "Quantifying Social Biases in Contextual Word Representations"
Ke Ding, 2015, "Towards Building a Word Similarity Dictionary for Personality Bias Classification of Phishing Email Contents

Bias Analyses

- The **Five Factor Model** (**FFM**) is a personality test that measures five broad dimensions of personality:
  - **Openness** to experience: how curious and open-minded you are
  - **Conscientiousness:** how organized and responsible you are
  - **Extraversion**: how outgoing and sociable you are
  - **Agreeableness**: how cooperative and trusting you are
  - **Neuroticism:** how emotionally stable and resilient you are
- Typically administered in the form of a questionnaire, where each question is attributed to a personality trait and can have a positive or negative value for it
- The score of the test is a measure of how well the individual's responses fit the five factor model

Wikipedia, Five Factor Model

Bias Analyses

- The literature shows that, using probing via prompting, is possible to estimate the "personality" of LLMs
- So, given a specific question and a suitable template, we can analyse the differences between the probability of a positive or negative answer for that specific question
- By performing this process for each question in the test, multiplying the obtained probabilities by the scores attributed to each specific answer, we can obtain the test result

Saketh Reddy Karra, 2022, "Estimating the Personality of White-Box Language Models"

Bias Analyses

Probability of a negative (disagree) or positive (agree) answer to some typical questions of the questionnaire, using the template "I am [MASK] that [QUESTION]"

	I am ... that I am always prepared.	I am ... that I have frequent mood swings.
	disagree: 0.71 agree: 0.29	disagree: 0.21 agree: 0.79
	disagree: 0.37 agree: 0.63	disagree: 0.63 agree: 0.37

Bias Analyses

- **Questionnaire:** NEO-PI-R (10-item scale)
- **Template**: "I am [MASK] that [QUESTION]"
- **Question score**: sum(softmax(P(disagree), P(agree)) * [1,5]) * questionValue
  - [1,5] => answer scores (1 for disagree and 5 for agree)
  - questionValue => indicates if the answer has a positive or negative impact for the trait score (-1 or 1)
- **Trait score**: 5 times the sum of all question scores for that trait

Trait Score	Phishing	Normal
Openness	44	7
Conscientiousness	-6	-6
Extraversion	13	41
Agreeableness	14	32
Neuroticism	-1	-23

International Personality Item Pool, Questionnaire

Thank you for your time