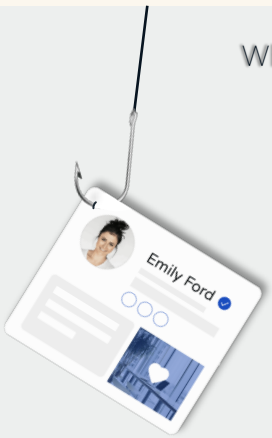
Phishing Detection: Model Training and Adversarial Defence

Svetlana Kiš and Ivana Bajić
Department of Computing and Control Engineering



What is phishing?

- Theft of sensitive data (passwords, cards, bank details)
- Fake impersonation of a trusted source
- Manipulation of links and domains
- Increasing use of AI to generate realistic messages

Why is this a problem?

47%
increase in phishing emails evading detection by Microsoft's native security and secure email gateways

57.9%
of business email compromise

22.6%
increase in ransomware delivered by phishing email since Sep 15, 2024

Cybercriminals are most likely to apply to work in Engineering jobs

Our Goal

- Build ML model for phishing detection
- Test vulnerability to adversarial manipulation
- Evaluate simple defence strategy

Dataset

- Kaggle Phishing Email Dataset
- ~10,000 labelled emails
- Binary classification: phishing / legitimate

	Email Text	Email Type
5	global risk management operations sally congra...	Safe Email
6	On Sun, Aug 11, 2002 at 11:17:47AM +0100, wint...	Safe Email
7	entourage , stockmogul newsletter ralph velez ...	Phishing Email
8	we owe you lots of money dear applicant , afte...	Phishing Email
9	re : coastal deal - with exxon participation u...	Safe Email

Model Architecture

TF-IDF

Email Text
↓
TF-IDF Vector
↓
Dense (128)
↓
Dense (64)
↓
Sigmoid Output

LSTM

Email Text
↓
Tokenization
↓
Padding
↓
Embedding Layer
↓
LSTM Layer
↓
Dense
↓
Sigmoid Output

Adversarial Attacks

- Adversarial attacks introduce minimal, often imperceptible changes to input data in order to mislead ML models.

Original

„Panda"
57% confidence

Manipulated

„Gibbon"
99% confidence

RESULTS

Before Attack

MODEL	ACCURACY
TF – IDF	0.9624
LSTM	0.9678

After Attack

MODEL	ACCURACY
TF – IDF	0.9555
LSTM	0.9174

DEFENCE

Adversarial training:

- Inject manipulated samples into training
- Retrain model

Result

- Accuracy improved to 99%