Manipulation Detection
Project Proposal
266 Final Project

Karl-Johan Westhoff
email kjwesthoff@berkeley.edu

UC Berkeley School of Information
MIDS Course 266 Summer 2025 Section 2 (Natalie Ahn)

1 Introduction

The human factor is increasing in relation to cyber attacks. The 2024 Verizon DBIR report [1].mentions that 68% of all cyber breaches involved the human element with phishing being a major contributor LLM's can now generate perfect and very convincing email text payloads, making current detection systems based on bag-of-words models ineffective
The purpose of this project is to develop a model that can detect if an email is manipulative and tries to make you take actions that are not in your best interest.

2 Datasets

2.1 Training Dataset

The model will be trained on a dataset generated by Wang et al. [2] which is based on 4000 labeled dialogues from films.

2.2 Evaluation Dataset

We will use data sets with labeled phishing emails in combination with results from previous models.

3 Methods

We will build an inference model that can detect manipulated emails based on a deep neural network with transformer architecture.

4 Evaluation

Our main interest is to investigate if the model can extend existing phishing detection systems by detecting manipulating language in the emails. We will to look at false negative results from previous models, to see if the detection of manipulative text captures emails that were previously missed.

References

[1] Verizon Business. *2024 Data Breach Investigations Report*. Tech. rep. Accessed: 2025-05-21. Verizon, 2024. URL: https://www.verizon.com/business/resources/reports/2024-dbir-data-breach-investigations-report.pdf.

[2] [Yuxin Wang, Ivory Yang ASD Saeed Hassanpour, and Soroush Vosoughi]. "MentalManip: A Dataset For Fine-grained Analysis of Mental Manipulation in Conversations". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 3747–3764. URL: https://aclanthology.org/2024.acl-long.206.