Manipulative Language Detection in LLM-Crafted Phishing Attacks

Karl-Johan Westhoff
email kjwesthoff@berkeley.edu
Neha Dhage
email neha$_d$*hage@ischool.berkeley.edu*

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

1 Introduction

The human factor remains central in cyber attacks. The 2024 Verizon DBIR report notes that 68% of breaches involve the human element, with phishing as a key contributor. With LLM tools, bad actors can now craft highly convincing phishing messages that evade traditional detection. This project investigates whether NLP models can detect manipulative language—specifically, text designed to influence actions not in the reader's best interest.

Machine learning (ML) models like Naive Bayes and basic neural networks are widely used to filter email traffic for spam (which is an abundant problem). However, they are often limited to detecting specific words or obvious patterns. Newer approaches combine lightweight ML filtering with resource-heavy NLP methods for cases that are not clearly categorized by simpler filtering. Since phishing often exploits human psychology through language, this study focuses on detecting manipulative language and whether such detection may improve defenses against phishing. Although the focus is on cybersecurity, manipulative language also appears in areas such as coercive or abusive communication, highlighting its broader relevance. Our approach first models manipulation using the "Mental Manip" dataset, then explores its potential for phishing detection.

2 Literature

Salloum et al. [3] provide an overview of current ML and NLP methods used for phishing detection, which forms the foundational context for this project. Suhaima et al. [4] trained models like BERT on spam data, whereas our focus will be on specifically detecting manipulative language. Wang et al. [2] created a data set that targets dialogue manipulation, which will serve as our primary training set. Al-Subaiey et al. have compiled a large corpus of emails in [6] from various datasets, under phishing specific email body texts; this will be used for attempts to detect phishing texts.

3 Datasets

Labeled data sets focused on manipulation are rare. Most of the research has come from psychology, which provides insight into the techniques used for manipulation. Most existing data sets suitable for NLP applications are concerned with hate speech and abusive language, which has been a hot topic in relation to online chat forums.

3.1 The MentalManip Dataset

Wang et al. [2] introduced the "MentalManip" dataset, published on hugging face[5]. The dataset is based on 4,000 fictional dialogues from online movie scripts. The data is labeled with a detailed manipulation taxonomy in three dimensions.

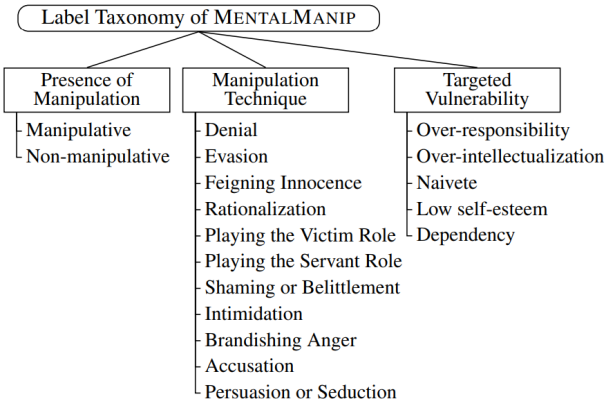

Figure 1: *Taxonomy*

3.2 Training Dataset

The model will be trained on a dataset generated by Wang et al. [1] which is based on 4000 labeled dialogues from films.

3.3 Evaluation Dataset

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

4 Methods

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

5 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] [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.