Manipulative Language Detection in LLM-Crafted Phishing Attacks

Karl-Johan Westhoff
email kjwesthoff@berkeley.edu
Neha Dhage
email neha_dhage@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 [1] 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. [2] provide an overview of current ML and NLP methods used for phishing detection, which forms the foundational context for this project. Suhaima et al. [3] trained models like BERT on spam data, whereas our focus will be on specifically detecting manipulative language. Wang et al. [4] 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 [5] 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 social media.

3.1 The MentalManip Dataset

Wang et al. [4] introduced the "MentalManip" dataset, published on hugging face [6]. The data set is based on fictional dialogues from "The Cornell Movie Dialogs Corpus" [7] from which suitable manipulative dialogues were selected using BERT and GPT-4 models, from these 4000 dialogues were manually selected to form the data set. The data is labeled with a detailed manipulation taxonomy in three dimensions see Figure 1, adding applied technique and psychological vulnerability mechanism to the binary presence of whether the dialogue contains manipulation or not.

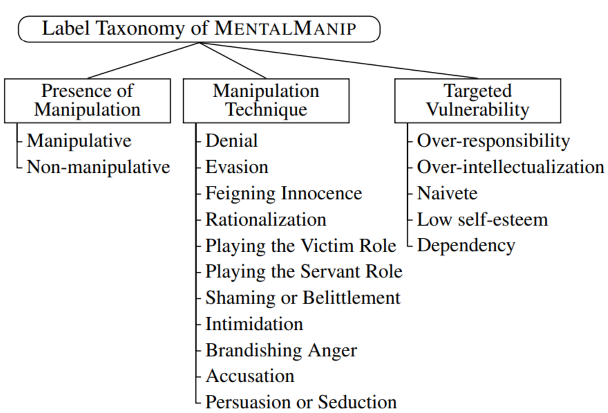

Figure 1: *Taxonomy of the MentalManip Dataset (labels in the data set) [4]*

The data set was manually labeled using a multi-phase human annotation process, adapting the taxonomy (see Figure 1) to the dialogue context three times by different people annotating. This gave two versions of the data set, one where the majority two out of three constitutes the result ("$MentalManip_{maj}$") and one where all three annotators have consensus and reach the same results ("$MentalManip_{con}$"). The $MentalManip_{maj}$ data set is larger (4000 rows) and more suitable for training a model capturing more instances of manipulation, the $MentalManip_{con}$ data set is smaller (2920 rows) and more precise and better suited for fine tuning. For this project we used the $MentalManip_{maj}$ data. In some cases these data fields are not complete in the data set requiring some degree of feature manipulation, This is addressed in section 4 below.

3.2 Data Exploration

The data set is not split equally between manipulation and non-manipulation, Figure 2 shows the distribution with 2.4 times more manipulation rows than non-manipulation (discussed in section 4).

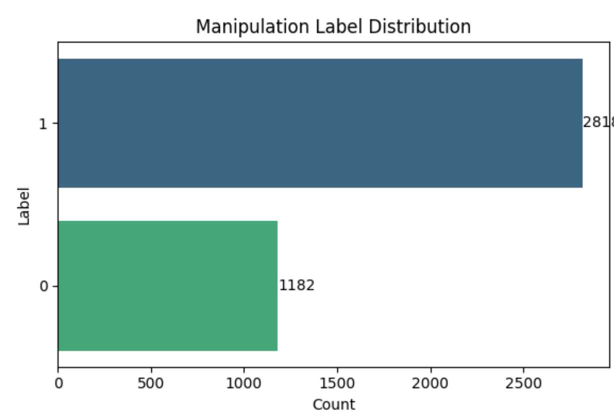

Figure 2: *Ratio of manipulation to non-manipulation in the $MentalManip_{maj}$ Dataset*

Some of the labels are missing for some of the rows with manipulation, Figure 3 shows 664[1] entries with missing technique labels,.

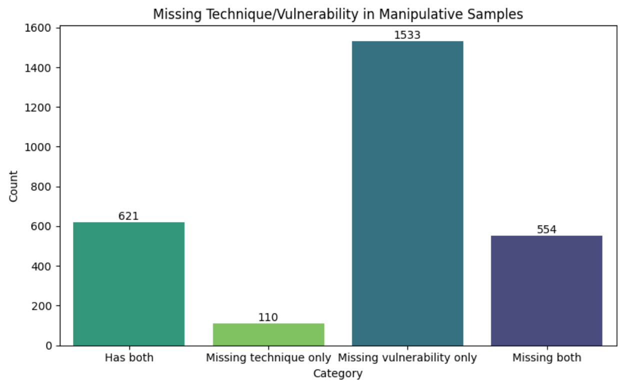

Figure 3: *Incomplete labeling of the MentalManip Dataset*

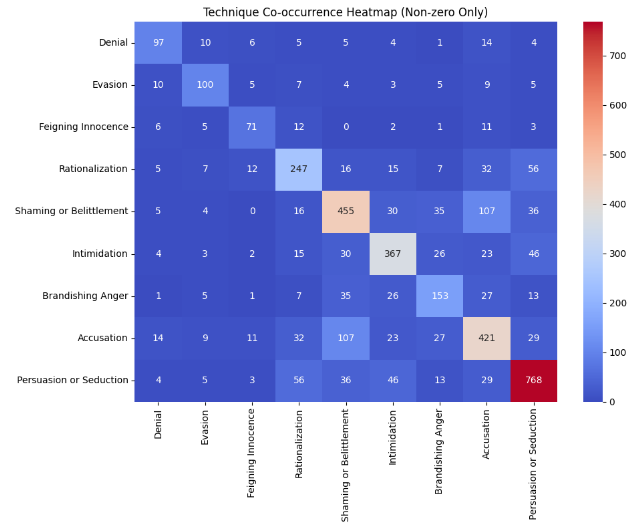

Figure 4: *Distribution and Co-ocurrence of technique labels*

Label Distribution The labels for technique and vulnerability mechanism are not uniformly distributed, furthermore, multiple labels occur in combination as comma separated values see Figure 4 showing "Persuasion or Seduction" as the most occurring for technique (this aligns well with use for phishing detection).

Further data exploration can be found in appendix A

4 Feature Engineering

- Address the ratio (e.g. use only the persuasion or seduction labels)
- Manipulate the labels - maybe merge the best technique and vulnerability mechanism that fits phishing
- Remove rows with missing text data
- etc..

We address the missing labels (see Figure 3) by either removing the rows with missing labels, or by imputing the missing values with an 'Other' category for the experiments with multi-label inference.

5 Experiments

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

6 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.

[1] 110 missing technique and 554 also missing vulnerability, we regard the technique labels as most relevant for phishing

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] Said Salloum et al. "Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey". In: *Procedia Computer Science* 189 (2021). AI in Computational Linguistics, pp. 19–28. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.05.077. URL: https://www.sciencedirect.com/science/article/pii/S1877050921011741.

[3] Suhaima Jamal, Hayden Wimmer, and Iqbal Sarker. *An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach*. Nov. 2023. DOI: 10.21203/rs.3.rs-3608294/v1.

[4] [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.

[5] Abdulla Al-Subaiey et al. *Novel Interpretable and Robust Web-based AI Platform for Phishing Email Detection*. 2024. arXiv: 2405.11619 [cs.LG]. URL: https://arxiv.org/abs/2405.11619.

[6] Yuxin Wang Ivory Yang Saeed Hassanpour 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://huggingface.co/datasets/audreyeleven/MentalManip.

[7] Cristian Danescu-Niculescu-Mizil and Lillian Lee. "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*. 2011.

A Data Exploration