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 aiming at 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.

4 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.

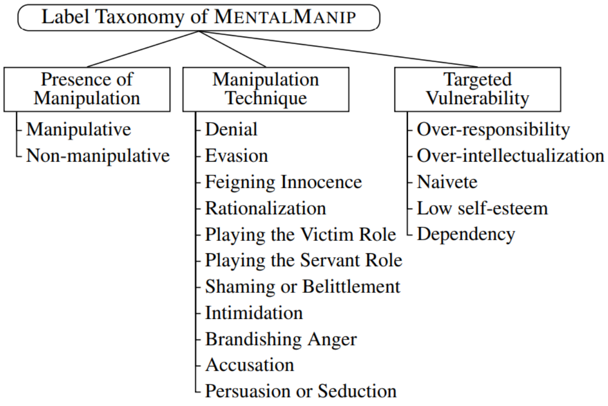

Figure 1: *Taxonomy labels in the data set*

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 6 below.

4.1 Data Exploration

4.2 Dialogues

The 4000 dialogues in the data set are between two persons exchanging sentences. By far the majority of dialogues consist of two exchanges, one by each person (there are only three cases with three exchanges). Word count statistics are shown in Figure 2, most dialogues consist of up to 50 words per person, and the number of words uttered by each person is fairly balanced, with person 2 saying slightly more words than person 1 in the up to 50 word majority case. Figure 3 shows the distribution of token counts for the dialogues in the data set, tokenized using BERT-base as reference. Only a minor number of dialogues exceed the BERT-base embedding size of 512 tokens.

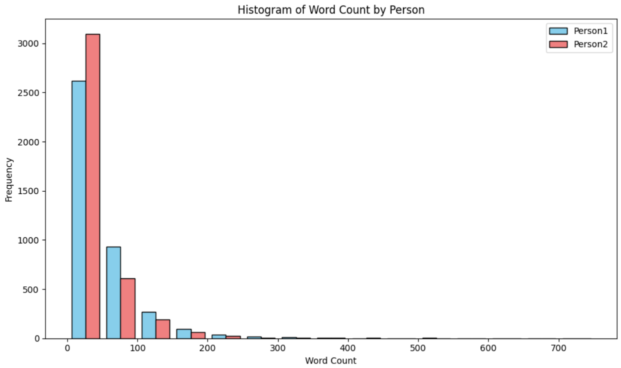

Figure 2: *Word count statistics for the dialogues in the MentalManip$_{maj}$ data set, words uttered by each person*

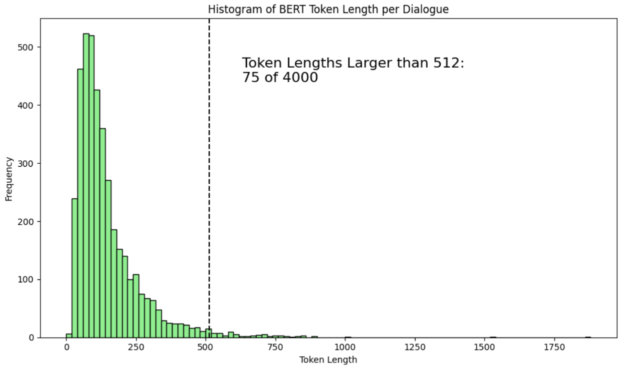

Figure 3: *Statistics for the dialogue in the MentalManip$_{maj}$ data set, tokenized using BERT-base*

4.3 Labels

Manipulation Label The data set is not split equally between manipulation and non-manipulation, Figure 4 shows the distribution with 2.4 times more manipulation rows than non-manipulation (discussed in section 6).

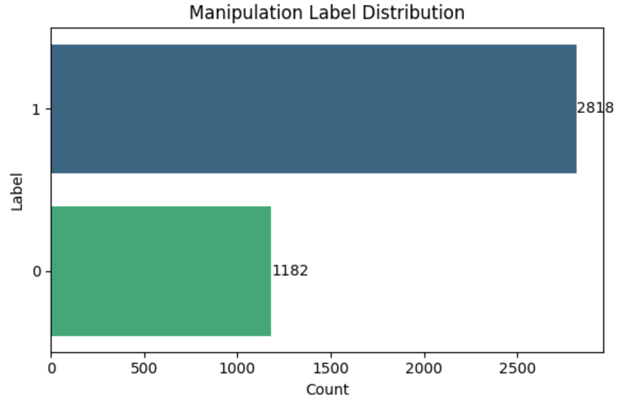

Figure 4: *Ratio of manipulation to non-manipulation in the MentalManip$_{maj}$ Dataset*

’Technique’ and ’Vulnerability’ Labels Some of the labels are missing for some of the rows with manipulation present[1], Figure 5 shows a total of 664[2] missing labels for ’technique’, we regard the technique labels as most relevant for phishing, especially the ’Persuasion or Seduction’ label.

The labels for ’technique’ and ’vulnerability’ mechanism are not uniformly distributed, furthermore, multiple labels occur in combination as comma separated values see Figure 6 showing co-occurrence for ’technique’ labels. "Persuasion or Seduction" is the most occurring ’technique’

[1] The labels should not be populated for non-manipulation rows

[2] 110 missing technique and 554 also missing vulnerability

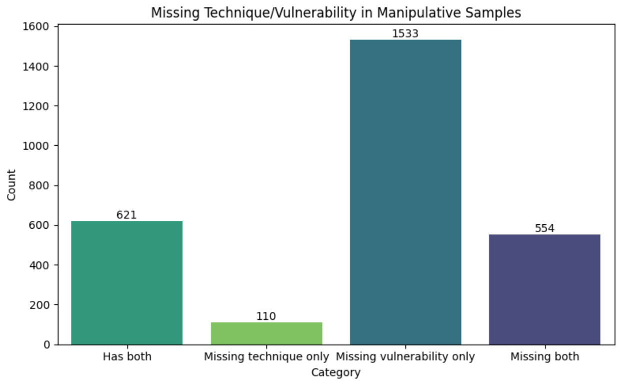

Figure 5: *Incomplete labeling of the MentalManip Dataset*

label (this aligns well with use with Phishing).

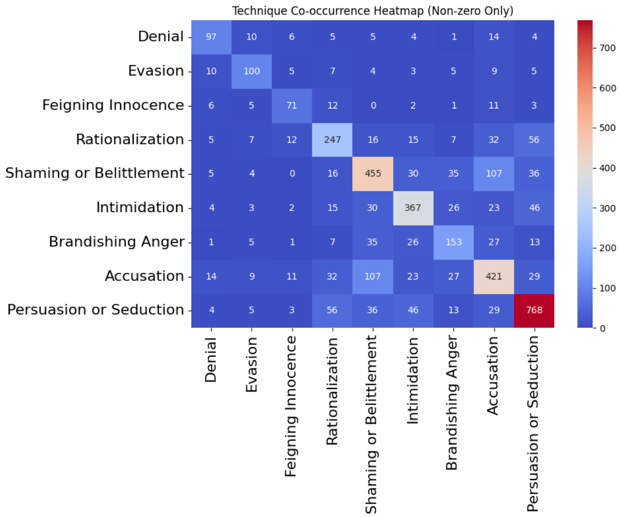

Figure 6: *Distribution and Co-ocurrence of technique labels*

Further data exploration can be found in appendix A

5 Baselines

With the relatively short embeddings (see Figure 3), the more basic versions of BERT have sufficient capacity to handle the data. The MentalManip article [4] also uses some decoder only models by 'zero' and 'few-shot' prompting the model with random example from the data set. This seems to perform better for overall binary classification, but only a little, and the LLM's have a tendency to pick up on toxicity and hate-speech and identify these as manipulation. Considering the label inconsistencies for 'technique' and 'vulnerability' in the data set, we will focus on the binary classification of manipulation for choosing a baseline model for further experimentation.

5.1 Binary with BERT and Buddies

Models looking at the 'manipulative' labels are trained on the $MentalManip_{maj}$ data set. The models were run with similar parameters, and the Accuracy at epoch before significant over fitting[3] recorded. Following models were investigated:

- BERT-base [8]
- RoBERTa [9]
- DistilBERT [10]
- ModernBERT [11]

Furthermore some "emotionally wiser" BERT derivatives exist which are pre-trained for emotion detection:

- BERTweet [12]
- EmotionBERT [13]

5.2 Baseline Results and Discussion

Results with losses and accuracy are shown in Table 1. The models were run until significant over-fitting occurred. In general the models over-fit after a few epochs

[3] Significant over fitting defined as: training loss / evaluation loss < 0.6

which is to be expected with a model that is extended from pre-trained. The Accuracy results are around 0.70-0.72 with little variation.

ModernBERT The model does not perform better, this was expected as the embedding lengths are short and not leveraging the benefits of ModernBERTs larger capacity.

Emotionally intelligent BERT BERTweet performs on par with BERT-base, the model is primarily trained for "Part-of-speech tagging", "Named-entity recognition" and "text classification" [12] herunder including emjojis etc. i.e. The model is not per-se expected to be better at manipulation detection, but we thought to give it a try.

Advanced BERT We also tried some more advanced BERT derivatives (DestilBERT and deBERTa_v3_small), however these models did not perform better than RoBERTa, and they required more compute resources to train. DeBERTa uses more advanced training loss, pre-training more advanced encoding etc.[14], however only the smallest version of deBERTa was possible to train with the hardware available. These models are optimized to deliver faster inference, but the extra cost in training resources make them less feasible for this project.

RoBERTa The best performing model reported in Table 1 was RoBERTa, which seems to perform slightly better than BERT-base, however with multiple tries the performance was not consistent, sometimes BERT performed better, however RoBERTa seemed more stable giving consistent results above 0.72 and performing more Epochs before over-fitting.

Model	**Epoch**	**Loss T/V**	**Acc Epoch**	**Acc Final**
BERT	2	0.97	0.726	0.70
roBERTa	4	1.03	0.728	0.73
deBERTa_v3	3	0.68	0.704	0.72
DistilBERT	2	0.75	0.709	0.72
ModernBERT	2	0.81	0.718	0.72
BERTweet	2	0.97	0.705	0.70
EmotionBERT	2	1.04	0.704	0.70

Table 1: *Base Model performance comparison across different transformer architectures for binary inference on the "manipulative" column: Epochs before significant over-fitting, Training loss / Validation loss (to measure overfitting) and accuracy at epoch and final classification*

6 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 5) 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.

7 Experiments

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

8 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] 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.

[8] Google AI. *BERT-base*. https://huggingface.co/google-bert/bert-base-uncased. Accessed: 2025-07-21. 2019.

[9] Facebook AI. *RoBERTa*. https://huggingface.co/FacebookAI/roberta-base. Accessed: 2025-07-21. 2019.

[10] Hugging Face. *DistilBERT*. https://huggingface.co/distilbert-base-uncased. Accessed: 2025-07-21. 2019.

[11] Answerdot AI. *ModernBERT*. https://huggingface.co/answerdotai/ModernBERT-base. Accessed: 2025-07-21. 2021.

[12] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. *BERTweet: A pre-trained language model for English Tweets*. 2020. arXiv: 2005.10200 [cs.CL]. URL: https://arxiv.org/abs/2005.10200.

[13] BorisN. *EmotionBERT*. https://huggingface.co/borisn70/bert-43-multilabel-emotion-detection. Accessed: 2025-07-21. 2023.

[14] Pengcheng He, Jianfeng Gao, and Weizhu Chen. *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. 2023. arXiv: 2111.09543 [cs.CL]. URL: https://arxiv.org/abs/2111.09543.

A Data Exploration

`https://drive.google.com/file/d/1s1mIXE58cj8miIoWOiB73VmZaSPs5vc5/view?usp=drive_link`