Detecting Psychological Techniques in Real-World Scams via Machine Learning

Allie Britton and Aryan Patel
Computer Science and Engineering Department, University of Notre Dame

UNIVERSITY OF NOTRE DAME

Background

Scams are becoming increasingly prevalent with scammers employing psychological techniques (PTs) such as authority, fear, and urgency to manipulate victims. Despite their widespread use these PTs have not been studied in the context of scams, making it difficult to identify or prevent scams effectively. Proper education around these PTs will help potential victims identify scam attempts before they are victimized.

Psychological Techniques

Authority and Impersonation
"Hello this is Emma from Amazon"

Fear and Intimidation
"You will be arrested, and your data will be leaked"

Urgency and Scarcity
"You must respond within two business days"

Pretext and Trust
The message contains accurate personal details.

Liking
"We are available 24/7 to help you in any matter"

Phantom Riches
"You were randomly selected to win $1,800,000"

Reciprocity
"We protect your account – now help us keep it secure."

Consistency
Scammer begins with a simple request to lead into the scam.

Social Proof
"Your resume was recommended by multiple recruiters."

Process

Data Annotation

We manually **labeled** a dataset of **700 real-world scam reports** sourced from reputable online platforms such as the Better Business Bureau (BBB) Scam Tracker. After identifying the PTs present in each message, we conducted two rounds of revision to ensure accurate labeling.

Data Preprocessing

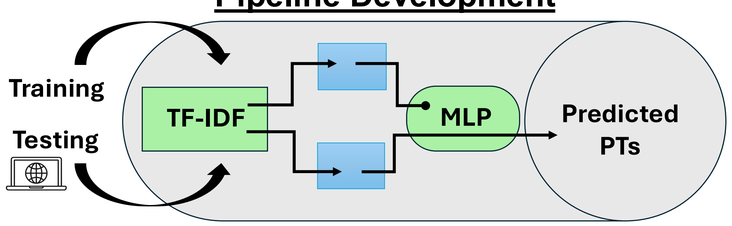
We applied standard **Natural Language Processing (NLP)** techniques to preprocess the labeled dataset. Techniques included **punctuation removal**, **stop-word removal**, and **tokenization,** as well as splitting the dataset into training and testing sets (80/20). Subsequently, the data was vectorized using two methods – **Term Frequency - Inverse Document Frequency (TF-IDF)** vectorization and **Word2Vec** embeddings – in order to evaluate their performance on our dataset.

Model Training and Validation

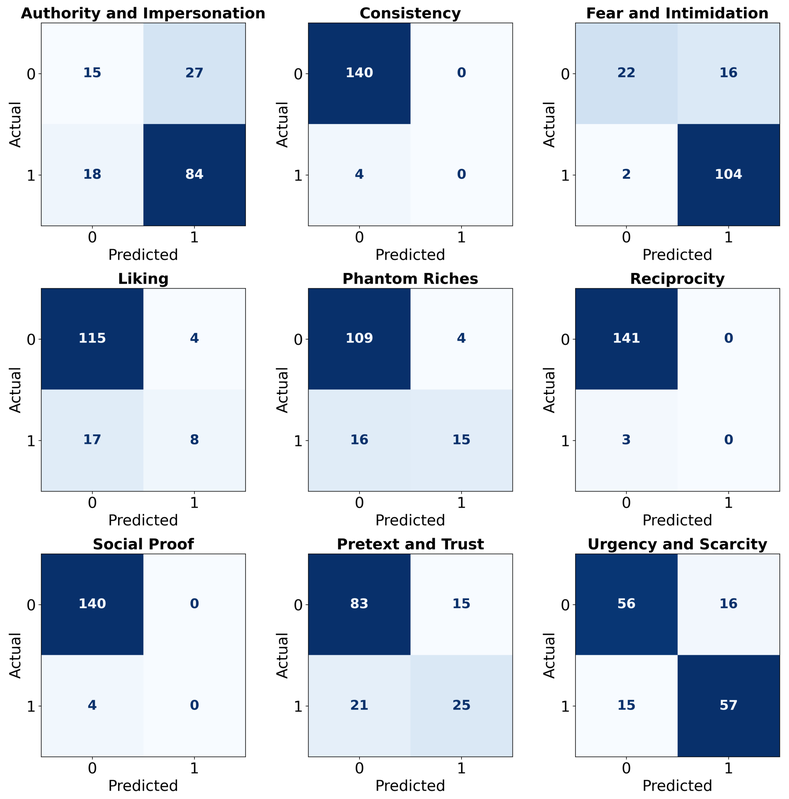
We employed Scikit-Learn's multi-class classification models, training each on the designated training set. Their performance was subsequently evaluated on the validation set and compared to determine the most suitable model for our task.

Model	Word2Vec Validation F1 Score	TF-IDF Validation F1 Score
Logistic Regression	0.627894	0.696923
Random Forest	0.712776	0.733024
Support Vector Machine (SVM)	0.616094	0.712334
K-Nearest Neighbor	0.703615	0.734879
Multilayer Perceptron (MLP) - Neural Networks	0.655328	0.751572

Pipeline Development



Model Results



GitHub Repository

Local Website

Conclusion

We develop a benchmark pipeline that is part of a full-stack web application designed to educate users on PTs utilized in scams. We expect the pipeline to produce better results upon the acquisition of more training data that contains scams with a better distribution of PTs utilized.

Acknowledgements

As part of the CSE-CSA Program, this work is partially sponsored by NSF under Grant No.CNS-2211428. We thank our mentor, Shang Ma, faculty advisor Yangfang Ye, and our program director Arturo Russell, for this research opportunity and their guidance.