Phishing email detection through linguistic patterns and sentiment analysis

Student: Daniel Nascimento Pedrinho

Supervisor: João Rafael Duarte de Almeida

Co-supervisors: Sérgio Guilherme Aleixo de Matos

Summary: With the wide usage of e-mail as a communication tool, phishing attacks have become increasingly common and sophisticated. This dissertation aims to explore **the use of linguistic patterns** and **sentiment analysis to detect phishing emails**. By analyzing the emotional tone and language used in emails, we may be able to identify potential phishing attempts and improve email security.

Motivation

- Email as Critical Infrastructure:
  - Over 4 billion users and 347 billion emails exchanged daily
  - Central to business operations, and online authentication
- Escalating Email Security Threats:
  - Phishing remains a prevalent attack vector
  - Attacks lead to data breaches and financial losses
- Increased Sophistication of Attacks:
  - Employs usage of social engineering, paired with AI generation
  - Emotional manipulation disregards technical prowess
- Novel Defense Approach:
  - Leverage emotion detection as a defense mechanism
  - Apply NLP and Machine Learning

State Of The Art

- Evolution of Detection Systems:
  - Rule-based and Blacklist Systems
  - Classical Machine Learning
  - Deep Learning and Transformer Models
- Trend Towards Multi-Modal Detection:
  - Integration of multiple email features, such as headers, in the detection
- Emerging Role of Sentiment Analysis:
  - Studies show sentiment-aware features can improve detection
- Dataset Landscape:
  - Large-scale datasets exist, providing binary labels
  - Fine-grained emotional datasets are mostly absent

Research Gap

- Phishing Remains Fundamentally Psychological:
  - Emotional manipulation is a primary success factor
- Lack of Fine-Grained Emotional Annotations:
  - No publicly available large-scale datasets with emotional labels
  - Imposes limit on supervised learning for emotional patterns
- Underexplored Role of Emotion Detection:
  - Often treated as a secondary feature
  - Remains peripheral in state-of-the-art systems

Dataset Requirements

- No existing datasets fit our research requirements
- **Objective**: Create sizeable, emotion-annotated phishing email dataset
- Dataset Characteristics:
  - Fine-grained emotional annotations
  - Realistic phishing language and structure
  - Balanced emotional category distribution
- Dataset should support:
  - Emotion-Aware Phishing Detection
  - Supervised machine learning and evaluation

Dataset Creation

- Methodology:
  - Use pre-existing curated dataset with 480 entires as baseline
  - Generated large dataset with 10,000 entries using LLM
  - Annotate dataset manually with human annotations

- Post-Processing:
  - Remove any generated text not part of the email
  - Replace placeholders with fake generic information

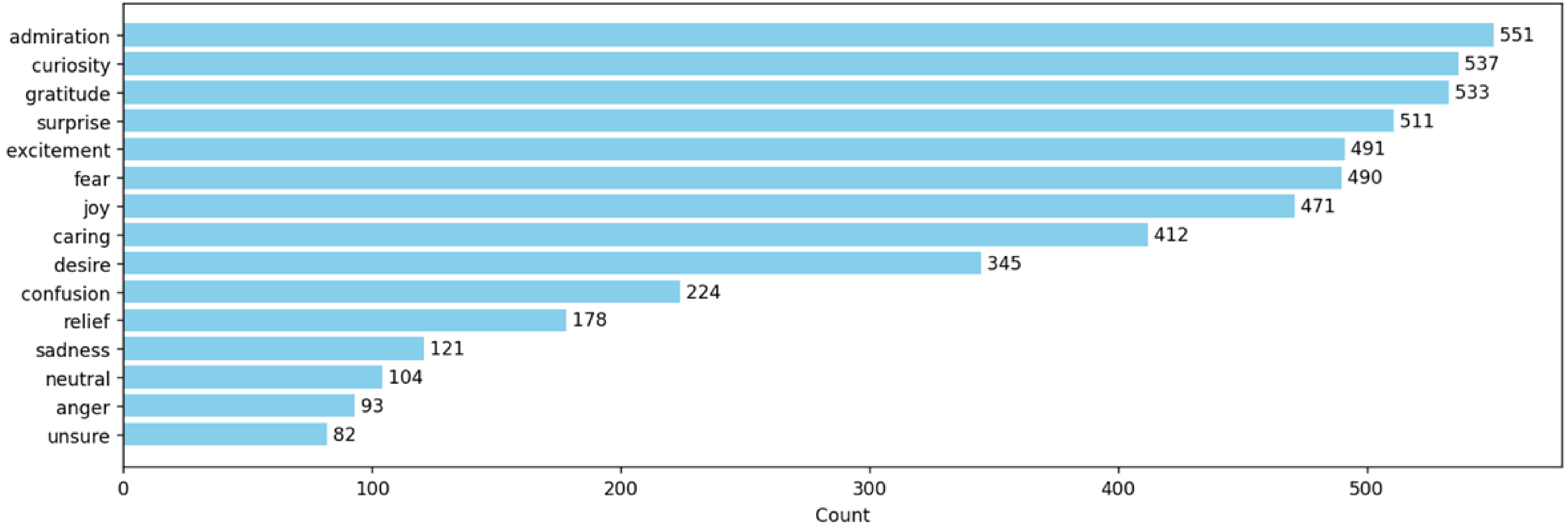
- Validation:
  - Dataset is evenly distributed across 14 emotions
  - Linguistic coherence preserved

Dataset Annotation

- **Objective**: Create Fully Annotated Dataset
  - Each entry can have more than one labe
  - Increases label diversity and training robustness
- Emotion labels derived from existing literature
- Annotation Procedure:
  - Emails are reviewed individually
  - Annotator chooses one or more emotions present in the email
- Quality Assurance:
  - Guidelines created to ensure consistency between annotators
  - “UNSURE” label prevents ambiguous labeling

Preliminary Results

- Dataset Creation Process Completed
- Annotation Process Underway:
  - 1843 entries annotated so far



Work Plan Proposal

- Phase 1: Phishing Detection Model Development
  - Model selection, training and evaluation
  - Parameter optimization
- Phase 2: Sentiment Analysis Model Development
  - Capable of identifying emotions with high scoring metrics
  - Integration with phishing model
- Phase 3: Web Application Development
  - Develop RESTful API that integrates models
  - Design frontend for user ease-of-use
- Phase 4: Review and Quality Assurance
  - Assure system is tested end-to-end
  - Review documentation and code aiming for high quality

Thank You!