Effect of AQUA Cleaning on Cross-Dataset Reliability in Phishing Email Detection

Caio Martini De Souza
cmartini@cougarnet.uh.edu

1. Problem & Hypothesis

- Phishing datasets contain inconsistent formatting, signatures, headers, and artifacts that cause models to overfit to dataset-specific noise instead of phishing-related content. AQUA cleaning should reduce this noise, improving cross-dataset reliability while causing minimal changes to within-dataset accuracy.

2. Datasets Used

Enron Email Dataset

- **Classes:** 2 (legitimate, phishing)
- **Class Sizes:** ~3,000 legitimate, ~1,500 phishing
- **Notes:** Real corporate emails; diverse writing styles; strong baseline.

Naser Phishing Email Dataset

- **Classes:** 2 (legitimate, phishing)
- **Class Sizes:** ~1,800 legitimate, ~1,784 phishing (balanced)
- **Notes:** Modern phishing emails; rich variety of malicious cues.

Twente Phishing Corpus

- **Classes:** 2 (legitimate, phishing)
- **Class Sizes:** ~264 phishing, ~300 legitimate (small + narrow)
- **Notes:** Template-like phishing; limited linguistic variety.

3. AQUA Cleaning Pipeline

- Lowercasing & normalization
- HTML removal
- Header & footer stripping
- Signature removal
- Deduplication
- URL / date / number normalization

4. Methodology

Models evaluated:

- Logistic Regression (TF-IDF baseline): Fast, interpretable, and sensitive to noise.
- DistilBERT transformer: Learns semantic content but disrupted by formatting inconsistencies.

Dataset versions:

- Raw datasets containing signatures, HTML, and formatting noise.
- AQUA-cleaned datasets with standardized and normalized text.

Evaluation approach:

- Cross-dataset testing: Train on Dataset A → Test on Dataset B.
- Within-dataset testing: Compare raw vs cleaned performance.
- Metrics: Accuracy, precision, recall, F1.

Purpose: Assess whether AQUA improves generalization, not just accuracy.

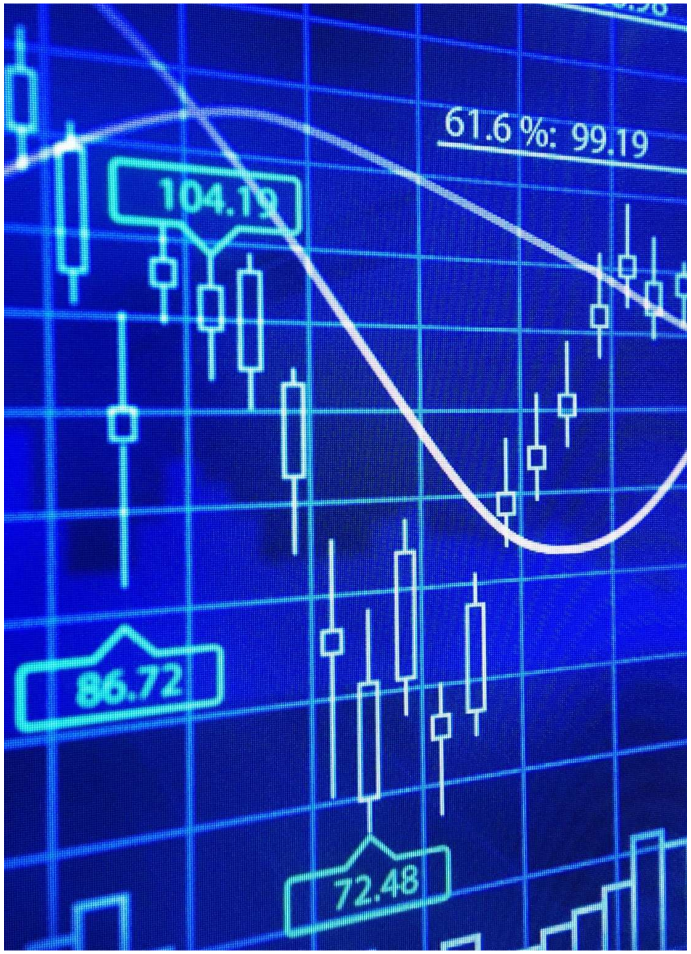
5. Key Results Summary

- Within-Dataset Accuracy Improvements:
- BERT – Enron: 0.9984 → 0.9998 (+0.14%)
- LR – Naser: 0.9820 → 0.9910 (+0.91%)

- Cross-Dataset Improvements:
- BERT – Enron → Naser: 0.67 → 0.83 (+0.16 absolute, +23.9%)
- LR – Enron → Twente: 0.55 → 0.72 (+0.17 absolute, +30.9%)

5. Key Results Summary

- AQUA improved cross-dataset **accuracy** for both DistilBERT and Logistic Regression, with the largest gains occurring when training and testing on datasets with highly different writing styles. Within-dataset performance changed only slightly for both models, indicating AQUA removes noise without altering meaningful content.
- DistilBERT still benefited the most from cleaning, showing strong improvements—especially in generalization—while Logistic Regression also showed substantial gains, including some of the largest absolute accuracy jumps (e.g., Enron → Twente).
- Twente-trained models continued to perform poorly after cleaning due to the dataset's limited diversity, which restricts generalization regardless of model type.
- Overall: AQUA significantly enhances stability and cross-dataset reliability for *both* models, improving real-world robustness.



6. Conclusion

- AQUA cleaning successfully removes dataset-specific noise that harms generalization.
- Results demonstrate that formatting artifacts—not linguistic differences—cause most cross-dataset failures.
- Transformers like BERT benefit strongly from cleaning because they encode structural patterns.
- Cleaning alone cannot overcome limitations of low-diversity datasets such as Twente.
- AQUA is an effective preprocessing tool for building robust phishing classification systems.

7. What I Learned

- How formatting noise and dataset inconsistencies can impact machine-learning outcomes.
- Why dataset diversity plays a critical role in cross-domain generalization.
- How AQUA cleaning enhances stability and reduces dependence on dataset-specific artifacts.
- Why transformer models benefit more from structured cleaning compared to linear models.
- The importance of cross-dataset evaluation for building reliable cybersecurity ML systems.

8. References

- 1. Champa, A. I., Rabbi, M. F., & Zibran, M. F. (2024). Curated datasets and feature analysis for phishing email detection with machine learning. IEEE ICMI.
- 2. Miltchev, R., Rangelov, D., & Evgeni, G. (2024). Phishing validation emails dataset. Zenodo.
- 3. Al-Subaiey, A., Al-Thani, M., Alam, N. A., Antora, K. F., Khandakar, A., & Zaman, S. A. U. (2024). Novel Interpretable and Robust Web-based AI Platform for Phishing Email Detection. arXiv preprint.
- 4. Enron Email Dataset. Miltchev, R. (Creator), Rangelov, D. (Creator), Evgeni, G. (Creator) (29 Aug 2024). Phishing validation emails dataset. Zenodo. 10.5281/zenodo.13474745
- 5. Al-Subaiey, A., Al-Thani, M., Alam, N. A., Antora, K. F., Khandakar, A., & Zaman, S. A. U. (2024, May 19). Novel Interpretable and Robust Web-based AI Platform for Phishing Email Detection.
- 6. A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Curated datasets and feature analysis for phishing email detection with machine learning," in *3rd IEEE International Conference on Computing and Machine Intelligence (ICMI)*, 2024, pp. 1–7.