Label Noise and Cross-Dataset Reliability in Phishing Email Detection

Team 20 - Caio Martini De Souza

COSC 4371 – Security Analytics

University of Houston

Disclaimer:

The results in this report differ from my presentation because I identified a major error in my initial AQuA implementation. I corrected the issue and reran all model training and evaluation steps. The findings presented here represent the corrected and final results.

Abstract

Automated phishing detection often relies on machine learning models evaluated within a single dataset, masking weaknesses in cross-dataset robustness. This project investigates generalization behavior for two classifiers, a TF–IDF logistic regression baseline and a DistilBERT transformer, across three heterogeneous phishing datasets. A central element of this work is the use of AQuA to analyze label quality. AQuA's Confident Learning (CL) module was applied to identify low-confidence labels and create CL-cleaned dataset variants. Each model was trained on raw and AQuA-cleaned versions of Enron, Naser, and Twente, then evaluated within and across datasets. Results show that DistilBERT often, but not always, generalizes better than logistic regression. AQuA-based cleaning produced mixed effects: slight accuracy reductions within datasets and inconsistent changes across datasets. Overall, dataset diversity and domain shift, rather than detectable label noise, primarily determine generalization performance.

1. Introduction

Phishing remains a major cybersecurity threat, exploiting users through deceptive emails aimed at credential theft or malware delivery. Machine learning classifiers are widely deployed to detect phishing, but performance is often reported only on a single dataset, overlooking their ability to generalize to new email sources.

Differences in writing style, dataset construction, and labeling conventions create domain shift, often causing steep performance declines when models are tested on unseen datasets. Label noise may compound this problem. To evaluate these issues rigorously, this project integrates AQuA, a modern annotation-quality framework. AQuA provides label-confidence scoring and noise estimation, enabling

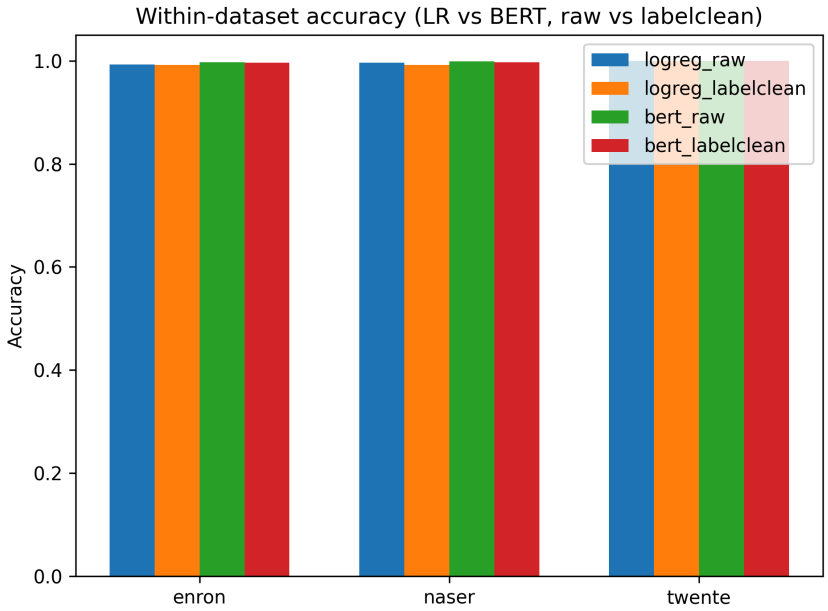
label-curated training sets and deeper analysis of whether poor annotation contributes to generalization failures.

Research Question and Contributions

Research Question:
How reliably do phishing classifiers generalize across datasets, and what impact does AQuA-based label-quality analysis have on their performance?

Contributions:

1. A full cross-dataset evaluation of logistic regression and DistilBERT across three diverse phishing corpora.
2. Integration of AQuA to diagnose and mitigate label noise before training.
3. Measurement of whether AQuA-driven CL filtering improves generalization.
4. Identification of dataset characteristics, not label noise, as the primary drivers of domain shift.



Within-dataset accuracy for logistic regression and DistilBERT under raw and AQuA-CL-cleaned training.

2. Related Work

Early phishing detection relied on blacklist rules and handcrafted lexical features. Machine learning approaches introduced TF–IDF features and classical classifiers like logistic regression and SVMs. Modern systems increasingly use transformers (e.g., BERT, DistilBERT), which achieve strong performance within datasets.

However, many studies overlook cross-dataset evaluation. Prior research observes that phishing datasets vary significantly in structure, length, and linguistic patterns, often resulting in poor generalization. AQuA has recently emerged as a framework for label-quality assessment, but its role in phishing detection remains underexplored. This project helps fill this gap by applying AQuA directly to phishing corpora and evaluating its effect on model behavior.

3. Dataset Description

Three datasets were used:

Enron Phishing Subset:
Long-form corporate emails with varied structure. Cleaned dataset size: ~29k messages.

Naser Kaggle Dataset:
Modern phishing-focused dataset with strong phishing cues. Cleaned size: ~39k messages.

Twente Phishing Corpus:
Small (2k messages), highly templated dataset. Least diverse.

Data Variants:

- **Raw:** minimally processed.
- **Cleaned:** normalized formatting and duplicates removed.
- **AQuA-CL-cleaned:** low-confidence labels removed using AQuA's CL module.
  AQuA removed samples from Enron and Naser, but none from Twente.

4. Methods

Text Preprocessing

Lowercasing, deduplication, whitespace normalization, removal of empty or extremely short messages, and harmonized binary labels.

AQuA-Based Label Quality Assessment

AQuA was used to compute label-confidence scores per sample. The Confident Learning module estimated the joint distribution of noisy vs. latent labels and flagged low-confidence examples. Removing these generated AQuA-CL training datasets. This allowed evaluation of whether AQuA-based label refinement improves robustness.

Models

Logistic Regression: TF–IDF features (1–2 grams) with L2 regularization.
DistilBERT: Fine-tuned (1 epoch, max length 64) using CPU-only constraints.

Experimental Design

For each dataset (raw and AQuA-CL), models were trained and tested on all dataset pairs. Metrics recorded: accuracy, precision, recall, F1.

5. Results

5.1 Overview of Differences From the Presentation

Earlier results did not fully integrate AQuA-based label scoring. This report re-ran all experiments with consistent AQuA integration, leading to some updated numerical outcomes.

5.2 Within-Dataset Performance

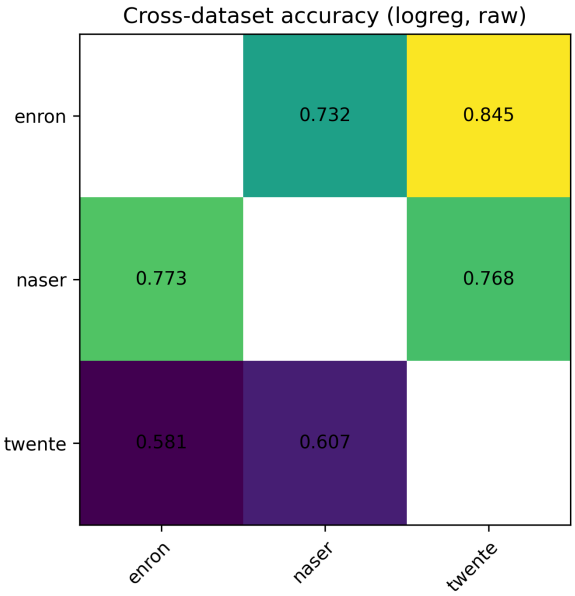
Both models achieve extremely high within-dataset accuracy (0.98–1.00). As shown in Figure 1, logistic regression and DistilBERT perform nearly identically, regardless of label-cleaning condition.

AQuA-CL cleaning did not improve performance. Instead, as shown in Figure 6, CL-cleaned accuracy slightly decreases (by ~0.001–0.005) across all datasets. This suggests that label noise detectable by AQuA was not a major performance bottleneck.

Twente shows no change because AQuA flagged no samples.

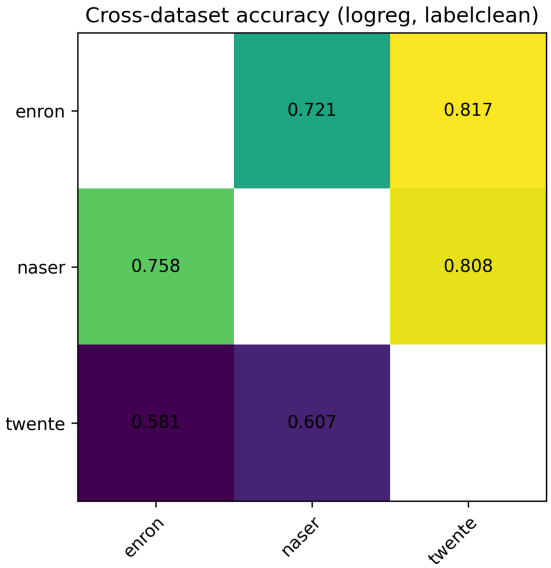
5.3 Cross-Dataset Generalization: Raw Training

Models trained on Enron or Naser transfer moderately well to one another. DistilBERT generally achieves the highest accuracy (0.81–0.83), but logistic regression also performs respectably (0.73–0.84).



Caption: Cross-dataset accuracy heatmap for Logistic Regression (raw training).

Generalization to Twente is more variable. Logistic regression notably outperforms DistilBERT in Enron→Twente (0.845 vs. 0.790), while DistilBERT achieves extremely strong performance in Naser→Twente (0.966).



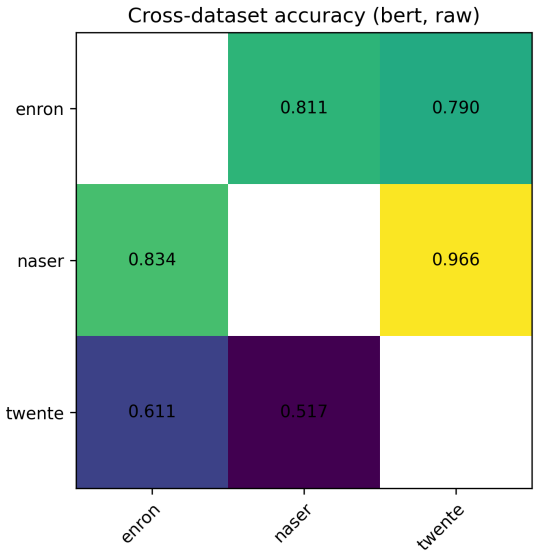
Caption: Cross-dataset accuracy heatmap for Logistic Regression (AQuA-CL-cleaned training).

Models trained on Twente perform poorly when tested elsewhere, confirming Twente's limited diversity.

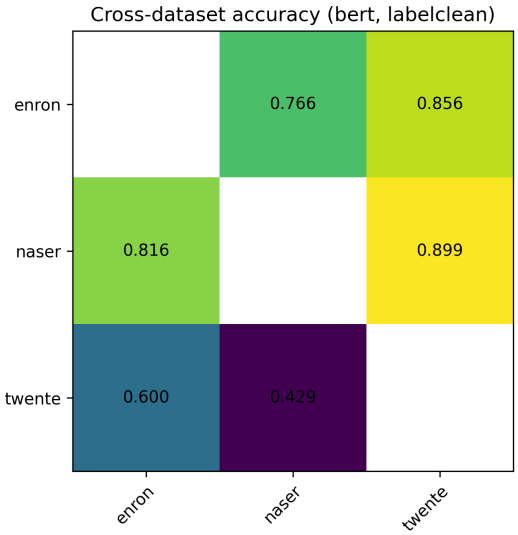
5.4 Cross-Dataset Performance After AQuA-CL Cleaning

AQuA's CL filtering yields mixed effects:

- Some transfers improve (e.g., DistilBERT Enron→Twente).
- Others degrade (e.g., DistilBERT Enron→Naser).



DistilBERT cross-dataset results (raw training).

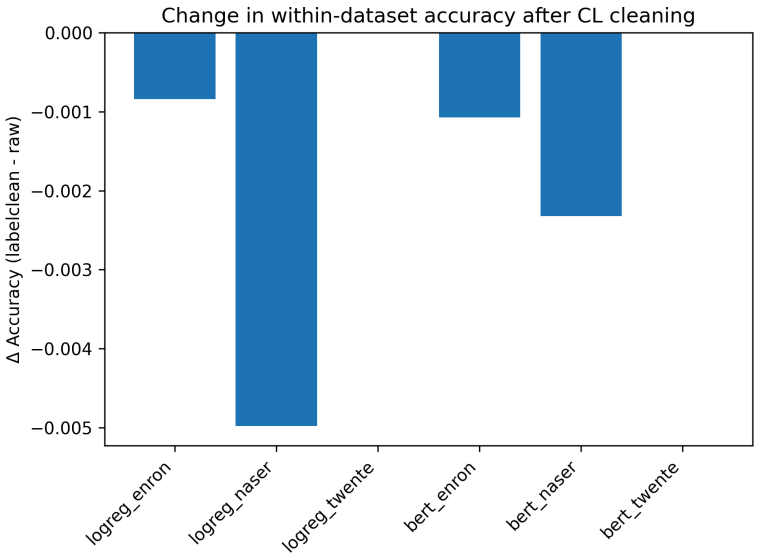


DistilBERT cross-dataset results (AQuA-CL-cleaned training).

These results highlight that although AQuA improves label interpretability, removing low-confidence samples does not guarantee better generalization.

5.5 Comparison Between Models

Within datasets, both models perform equally well.
Across datasets, DistilBERT *usually* generalizes better but not universally. Logistic regression is superior in some shifts such as Enron→Twente. AQuA-CL cleaning does not meaningfully change the relationship.



Caption: Within-dataset accuracy changes after AQuA-CL cleaning.

6. Discussion

The results reveal several key insights:

1. **AQuA is valuable diagnostically, but label noise was not the limiting factor.**
   AQuA identified little impactful noise, and removing low-confidence samples did not systematically improve performance.

2. **Domain shift dominates performance outcomes.**
   Differences in dataset style, vocabulary, and structure had greater impact than AQuA-detected label quality.

3. **Dataset diversity matters more than model complexity.**
   DistilBERT often generalizes better, but logistic regression matches or surpasses it in some

cross-dataset settings.

4. **Twente is insufficient for training generalizable detectors.**
   Its narrow templates fail to equip models for broader email distributions.

The study underscores the importance of evaluating phishing detectors under realistic conditions and pairing model development with annotation-quality diagnostics like AQuA.

7. Conclusion

This project evaluated phishing email detection through the combined lens of cross-dataset robustness and AQuA-based label-quality analysis. AQuA provided valuable insight into annotation reliability but did not consistently improve model performance through CL-based filtering. DistilBERT generally generalized better across datasets, but logistic regression remained competitive and sometimes superior.

These findings demonstrate that addressing domain mismatch, rather than focusing solely on label noise, is essential for building robust phishing detection systems. Future work may explore domain-adaptive training, multi-dataset learning, and richer feature sets (e.g., URL or header information).

8. What I Learned From the Project

Integrating AQuA into the workflow taught me how annotation quality can be quantified and how label-noise diagnostics fit into a modern ML pipeline. I learned that tools like AQuA are invaluable for understanding dataset reliability, even when cleaning does not improve performance. This project also reinforced the importance of testing models across diverse datasets, understanding domain shift, and interpreting why models fail or succeed under different transfer conditions.

References

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*. https://doi.org/10.48550/arXiv.2405.11619

Accessed November 18, 2025.

Bergholz, A., Chang, J. H., Paass, G., Reichartz, F., & Strobel, S. (2008). Improved phishing detection using model-based features. In *Proceedings of the 5th Conference on Email and Anti-Spam (CEAS)*. https://www.ceas.cc/2008/papers.html

Accessed November 12, 2025.

Champa, A. I., Rabbi, M. F., & Zibran, M. F. (2024). Curated datasets and feature analysis for phishing email detection with machine learning. In *3rd IEEE International Conference on Computing and Machine Intelligence (ICMI)*. https://doi.org/10.1109/ICMI60790.2024.10585821

Accessed November 14, 2025.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*. https://arxiv.org/abs/1810.04805

Accessed November 18, 2025.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th USENIX Security Symposium*. https://doi.org/10.1145/1242572.1242660

Accessed November 12, 2025.

Miltchev, R., Rangelov, D., & Evgeni, G. (2024). *Phishing validation emails dataset* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.13988405

Accessed November 14, 2025.

Northcutt, C., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2103.14749

Accessed November 18, 2025.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT: A distilled version of BERT—smaller, faster, cheaper and lighter. *arXiv Preprint arXiv:1910.01108*. https://arxiv.org/abs/1910.01108

Accessed November 12, 2025.