Behavioral Cues and Adversarial Robustness in Phishing Email Detection

Team Capybara

Ailina Aniwan | Jay Liu | Eric Ortega Rodriguez | Tursunai Turumbekova

Why Phishing Detection Needs a New Approach

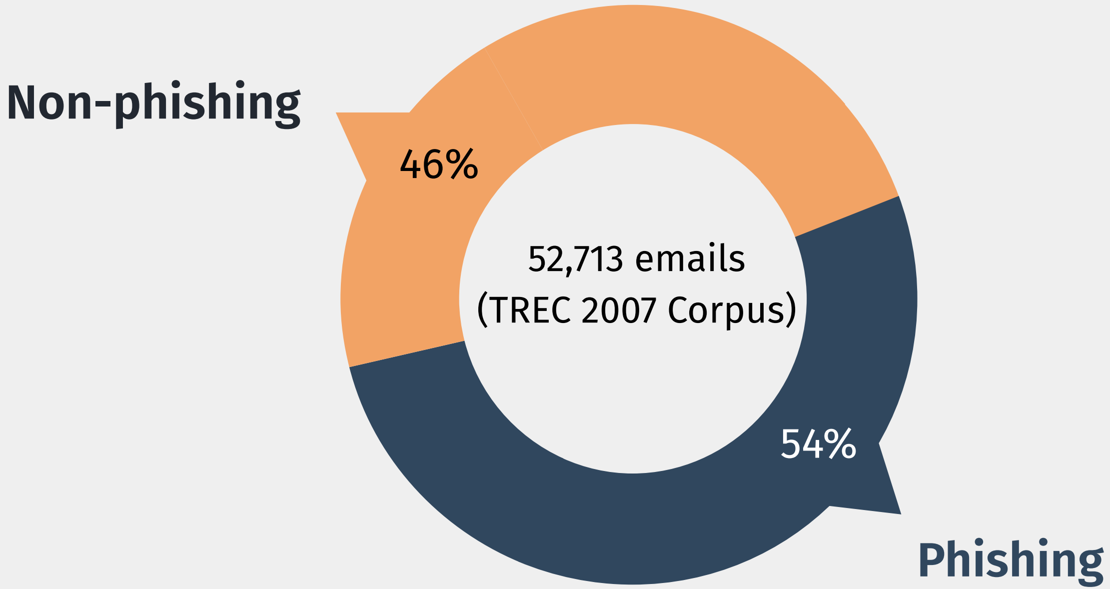
91% of cyberattacks begin with phishing
Deloitte (n.d.)

Phishing costs businesses over $4.91 million per breach
IBM Security Cost of a Data Breach Report (2023)

90% of phishing attacks exploit human psychology — urgency, fear, curiosity
(Microsoft, 2023)

Data

Rich features: sender, receiver, subject, body, URL counts
Why it matters: Enabled both text-only and metadata-enhanced modeling



Our Experimental Roadmap

01

Model Comparison: Naïve Bayes, Logistic Regression, XGBoost, BERT

02

Text Preprocessing: do standard preprocessing steps improve performance?

03

Adversarial Robustness: How do models hold up under small text corruptions?

04

Behavioral Features: Can psychological cues improve detection?

Experiment 1:
Which Model Wins?

Approach:

→ Compared: Naïve Bayes, Logistic Regression, XGBoost, BERT
→ Train only on email text, then on text + metadata

Result:

→ BERT: Best on text-only (99.68%)
→ XGBoost: Best overall with metadata (99.75%)

💡 **Hybrid features (text + metadata) beat pure text alone.**

Experiment 2:
Does Text Cleaning Help?

Approach:

→ Test: lowercasing, stopword removal, HTML stripping, punctuation splitting

Result:

→ Naïve Bayes: Stopword removal improved accuracy and reduced false positives
→ XGBoost & Logistic Regression: Stable — preprocessing had minimal effect

💡 **Preprocessing helps only in simpler models and can sometimes hurt.**

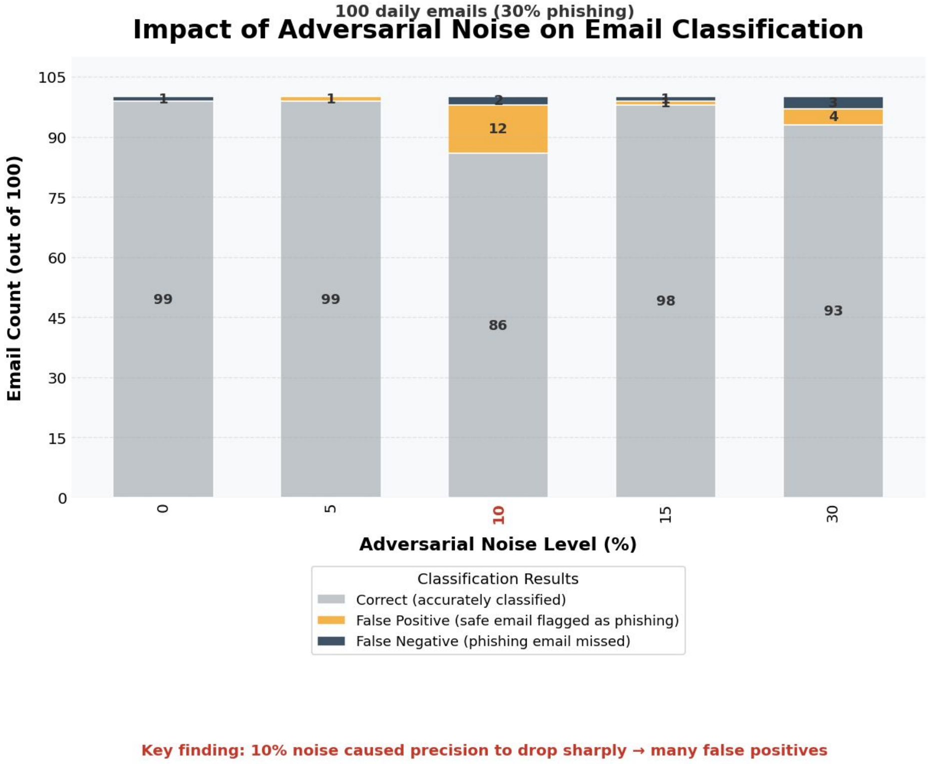
Experiment 3 – Can BERT Handle Noisy Attacks?

Approach:

- Injected synthetic noise at levels: 0%, 5%, 10%, 15%, 30%
- Evaluated accuracy, precision, recall, F1 across noise levels
- Simulated a real-world inbox: 100 emails/day, 30% phishing

Results:

- At 10% noise, F1-score dropped from 0.97 → 0.80
- Precision plummeted to 0.694 → 12 safe emails flagged as phishing
- Performance briefly rebounded at 15%, then declined again at 30%



Even small amounts of adversarial noise can severely degrade BERT's reliability

Experiment 4 – Detecting Psychological Manipulation in Emails

Imagine getting an email that says:
“Act now or your account will be closed.”

Approach:

- We created a dictionary of manipulative phrases — ones that scammers often use.
- Then we used Sentence-BERT to measure how similar each email was to those phrases.
- We added those deception scores into our models as new features.

Results:

- Adding behavioral scores improved Naïve Bayes and XGBoost
- XGBoost became 33% better at catching real phishing emails it used to let through

Takeaways

1. **BERT is smart, but not invincible**
   Performed great—until we added a little noise. A few typos, and it started making big mistakes.

2. **Simple tricks still work**
   XGBoost did great—when we gave them smart features like sender info and URLs. Cheap, fast, and effective.

3. **Phishing is psychological**
   We added scores for urgency, fear, and curiosity. That helped our models catch trickier emails—even without fancy AI.

4. **Every small gain matters**
   A 0.1% boost in accuracy may sound small, but at scale? That’s hundreds or thousands of threats stopped

Thank you! ☝🏻 🤓

We are ready to answer your questions

References

Abobor, M., & Josyula, D. P. (2023). SOCIALBERT: A transformer-based model used for detection of social engineering characteristics. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 174–178). IEEE. https://doi.org/10.1109/CSCI62032.2023.00033

Deloitte. (n.d.). 91% of all cyber attacks begin with a phishing email to an unexpected victim. *Deloitte Malaysia*. Retrieved March 20, 2025, from https://www2.deloitte.com/my/en/pages/risk/articles/91-percent-of-all-cyber-attacks-begin-with-a-phishing-email-to-an-unexpected-victim.html

IBM Security. (2024). X-Force Threat Intelligence Index 2024. IBM Corporation. https://www.ibm.com/reports/threat-intelligence

Microsoft Threat Intelligence. *Microsoft Digital Defense Report: Building and Improving Cyber Resilience*. Oct. 2023.

Shahriar, S., Mukherjee, A., & Gnawali, O. (2023). Improving phishing detection via psychological trait scoring. *arXiv preprint*. arXiv:2305.13263. https://arxiv.org/abs/2305.13263

Appendix: Ethical Considerations

Bias:

We trained our models on English emails from 2007. They might fail on newer or non-English phishing tactics, risking unfair protection gaps.

Risk of Misclassification

Behavioral features like urgency or fear are powerful—but may misflag legitimate emotional messages, especially from user groups with differing communication norms.

Privacy in Real-World Use

While our dataset is anonymized, real deployment means scanning live emails, raising concerns around surveillance, consent, and data governance. Strong encryption, clear retention policies, and auditability are essential safeguards.

Model Explainability

In high-stakes environments like legal or enterprise settings, users must understand why an email was flagged. We used interpretable models (Logistic Regression, XGBoost) and behavioral feature visualization to build trust.