Detecting AI-Generated Phishing Through Sender-Style Communication Drift

James Wilson III
CIS Department, Fordham University
New York, NY, USA
jiw2@fordham.edu

Nancy Martinez
CIS Department, Fordham University
New York, NY, USA
nm82@fordham.edu

***Abstract*—Phishing is a threat to organizational communication systems which will continue to exist; and the increasing use of Large Language Models (LLMs) in generating phishing emails has increased their credibility by enabling attackers to replicate the tone, structure and vocabulary used by employees within organizations. The goal of this study was to investigate whether AI-created phishing-type emails could be identified via sender-stylistic communication drift alone or if they would need to rely on the existing phishing indicator-based detection methods. Our anomaly detection setup showed communication drift may provide an effective secondary warning signal. These results indicate that using sender-specific stylometry may help reinforce organizational defenses against phishing attacks as an additional behavioral detection method.**

***Index Terms*—phishing detection, AI-generated, stylometric analysis**

I. Introduction

Phishing attacks through email-based communication systems are one of the biggest threats to an organization. Attackers can use trusted communications channels to exploit human behavior to reveal sensitive information. In organizations, employees constantly send and receive messages through email and other platforms. If a phishing attack were to be successful, it could lead to financial loss, data breaches, and unauthorized access to internal systems. As major organizational communications are digital, it is important to protect these channels.

Artificial intelligence has been on the rise, and so has the sophistication of phishing attacks. These deceptive emails are now being generated using artificial intelligence, increasing their ability to imitate legitimate corporate communication. This includes tone, structure, and vocabulary. These messages are designed to blend in with normal communications, unlike traditional phishing attacks that include suspicious formatting and grammatical errors. AI-generated emails are difficult to detect, as they can easily adapt writing styles

As the threat landscape evolves, traditional phishing detection methods, such as keyword-based filtering, are no longer sufficient. This is due to AI-generated emails replicating style and content. This creates a need for detection techniques that look into how messages are composed rather than what it contains. It is essential to analyze sender-style inconsistencies that evaluate incoming messages based on how much it deviates from a sender's historical communication behavior. This communication drift can provide signals to identify suspicious emails.

This research investigates the use of sender-relative communication drift detection to identify AI-generated phishing emails. By using the Enron email dataset and synthetically generated anomalous emails, this study models sender-specific writing styles with stylometric features and evaluates how deviations from these patterns can be used for detection. Stylometry is the statistical analysis of literary style, primarily used to determine the authorship of an anonymous or disputed document. The main focus of this work is the introduction and evaluation of a driftscore-based approach that measures stylistic deviation relative to a sender's baseline, demonstrating its potential in phishing detection systems.

II. Related Works

Traditionally, phishing detection methods have relied on rule-based filtering and machine learning techniques. They typically analyze features such as email headers, domains, keywords, and embedded URLS. These are usually used by Google (Gmail) and Microsoft (Outlook), which have spam filters that attempt to block malicious communications before they reach users. Research has shown that these systems cannot defend against sophisticated attacks. Especially threats that use advanced

language models that don't have obvious indicators of malicious intent[2].

Stylometry, defined as the statistical analysis of writing style to determine authorship, has been used as an alternative approach to find anomalies in communication. The patent US11516223B2 talks about a system that builds sender-specific writing profiles that use linguistic and behavioral features such as vocabulary usage, sentence structure, and punctuation patterns [3]. Deviations can be found by comparing incoming messages that can show malicious intent or impersonation. Research on the stylometric detection of AI-generated phishing emails shows that sentence length, writing consistency, and lexical diversity can distinguish between legitimate and engineered emails [2].

Studies have shown how AI-generated phishing attacks have increased in effectiveness. Industry reports revealed that 83 percent of phishing emails were AI-generated. Further demonstrating how accessible AI technologies are [1]. Works published by arXiv mention that large language models can personalize and create content-aware emails that mimic organizational communication style [5]. ScienceDirect confirms that these generated emails can bypass traditional detection systems due to their adaptability and sophistication [4], [6]. This research has confirmed how limited existing phishing defenses are.

Although there are plenty of defense mechanisms for phishing emails, there is a lack of research focusing on sender-relative communication drift as a defense. Existing work has looked into stylometric profiling and AI-generated phishing independently, but has yet to dive deep into measuring how an email deviates from a sender's historical communication patterns over time. This research builds on previous stylometric and anomaly detection research by introducing a DriftScore-based framework, which assesses these deviations and uses them to identify AI-generated emails within organizational communications systems.

III. Methodology

A. Dataset

This project utilized the Enron Corporate Email Dataset as a source of legitimate corporate communications. The dataset is composed of many emails from actual employees within the Enron organization from 1999 to 2003. Since the most updated dataset is saved in the 'maildir' file format as .eml files, the emails had to be converted into a structured CSV format. Additionally, to allow for adequate performance of stylometric analysis of the remaining emails for each sender, additional sender-based sampling was conducted so that each sender would have a minimum number of legitimate emails to evaluate.

To enable sender-based stylometry, the preprocessing pipeline was limited to including only those rows having a valid identifier and email body. Additionally, sender normalization was performed to standardize improperly formatted email addresses before analysis.

For the purpose of generating a comparative dataset for evaluating the effects of communication-drift; synthetic anomalous emails were created based upon sender-style profiles. Each sender received a legitimate email-based style profile utilizing aggregated stylistic characteristics such as average word length, average sentence length, punctuation behavior, lexical preferences, and representative examples. For testing, these profiles were saved in a JSON format that saved each sender's email count, average words per email, median words per email, average sentence length in words, vocabulary diversity, use of punctuation, uppercase ratio, and an array of their most used words besides stop words (e.g., the, a, an, and, etc.). A language model (gpt-4o-mini) was then called via an API using the official OpenAI Python Library to generate "defanged" anomalous business emails which would preserve the sender's writing style but contain obvious anomalies. Defanging is the delibrate process of modifying IOCs, and in this case, artifacts within suspicious/phishing-like emails.

The generation pipeline was intentionally constructed to generate secure research artifacts (i.e., not actual phishing content for ethical purposes). After preprocessing and filtering; there remained 11,106 legitimate Enron emails related to senders whose style profiles were retained out of the original ˜500k. Also remaining after filtering were 1,128 synthetic anomalous emails that were produced within the validated anomaly set. Due to a validation process that removed generations not meeting style-constraints for their respective senders; the number of synthetic anomalies remaining was less than originally desired since the goal was 2000. However, this smaller than anticipated synthetic anomaly set was deemed sufficient for proof-of-concept purposes.

B. Stylometric Feature Extraction

Each email was translated into a vector representing its stylometric characteristics. Attributes extracted from each communication included general writing attributes (e.g. word count, character count, sentence count, av-

erage length of sentences, average length of words, type/token ratio, etc.), as well as other linguistic and behavioral indicators (such as use of question mark and/or exclamation mark, comma, semicolon, presence/absence of greeting/sign-off flags, counts of modal verbs, counts of urgency cues and financial action cues, etc.) as well as ratios of use of pronouns and counts of imperative verbs, and counts of "defanged link" indicators such as "hxxps://" versus "https://".

In addition to extracting global stylometric attributes for each communication, attributes were also derived to measure a sender's level of deviation from his/her own writing style. This included attributes such as difference in word count compared to sender averages; difference in average length of sentences compared to sender averages; difference in punctuation density compared to sender averages; difference in upper case ratio compared to sender averages; and content-word overlap with sender-specific high frequency vocabulary. These attributes were developed to identify when a communication appears unusual given a sender's past writing behaviors versus merely identifying generic phishing signals.

C. Experimental Design

Two experiments were conducted. The first one investigated supervised stylometric classification. Legitimate synthetic anomalous emails were balanced by sender and then split into training/test sets. The baseline stylometric features of each sender were calculated using only legitimate training emails to prevent leakage. Both Logistic Regression and Random Forest models were trained to distinguish between legitimate emails and synthetic anomalous emails based upon the stylometric and drift related features. This experiment was used as a baseline to determine if there were enough discriminatory signals present in the feature space under labeled conditions. This supervised experiment was intentionally designed to validate the usefulness of the feature set before moving on to the more realistic and more difficult sender-specific anomaly detection experiment.

The second experiment investigated into the main purpose of this project, the ability to detect anomalies at the sender-specific level. For each sender, legitimate emails were divided into training, validation and test sets. Isolation Forest and One Class SVM models were trained using only the legitimate training emails. The validation set combined with a subset of synthetic anomalous emails was used to determine optimal sender-specific threshold values for detecting anomalies. A variety of decision modes were considered when developing the lightweight warning mechanism similar to those found in modern-day email clients. These included anomaly-model-only detection, drift-score-only detection and two-stage warning rules. In the two-stage setting a message was first identified as being out of bounds if its anomaly score or 'DriftScore' surpassed a tuned sender-specific threshold. A final warning would be issued if that departure from normal writing behavior was accompanied by at least One supporting suspicious cues such as urgent language usage, financial action language usage, imperative phraseology or low content-word overlap with the senders baseline. This design was implemented to reduce false positives since it ensured that the senders writing style deviated alone did not provide reason enough to issue a warning against suspect messages. For interpretability, we define a sender-specific DriftScore as a measure of how strongly an email deviates from the sender's specific writing style.

$$\text{DriftScore} = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{x_i - \mu_i}{\sigma_i + \epsilon}\right| \tag{1}$$

where x_i is the feature value for the evaluated email, μ_i and σ_i are the mean and standard deviation of that feature estimated from the sender's legitimate training emails, k is the number of selected features, and ϵ is a small constant used for numerical stability. Larger DriftScore values indicate that an email is more unusual relative to the sender's historical writing style.

IV. Results

A. Classification Performance

The supervised learning task proved that stylometric and sender-relative drift features contains a considerable amount of discriminative information even when all instances are labeled. The logistic regression model achieves an accuracy of 87.%, precision of 0.851, recall of 0.906, F1-score of 0.878, and ROC-AUC of 0.945. The random forest classifier demonstrates the greatest performance across all metrics with an accuracy of 93.9%, precision of 0.942, recall of 0.935, F1-score of 0.939, and ROC-AUC of 0.985. These values clearly show that legitimate an anomalous emails are very easily distinguishable when the two are present for training. In addition, the confusion matrix provides additional support for the findings above. For example, in Figure 1, we can see that logistic regression has 33 false positives and 15 false negatives. Random forest decreases those numbers by having 10 false positives and 12 false

negatives. Therefore, experiment 1 demonstrates that the extracted stylometric and sender-relative features were very effective in a supervised setting and provided a strong baseline for comparison against sender-specific anomaly detection.

B. Anomalous Sender-Specific Email Classification Performance

Unlike the supervised learner, the sender specific anomaly detection was significantly more challenging. When using Isolation Forest as the baseline anomaly detection method, it produces a precision of 0.24, a recall of 0.74 and an F-1 score of 0.36. Similarly, One-Class SVM also produced a lower precision of 0.27 and a lower recall of 0.63 along with a low F1 score of 0.38. The DriftScore as a singular metric performed similarly well, producing a precision of 0.23, a recall of 0.77 and an F-1 score of 0.36. It appears as though the sender style variability produces some discrimination; however, it does not appear sufficient to produce clean separation between legitimate and anomalous synthetic emails.

To provide a greater approximation to a user facing warning system within an email client, three different types of warnings (permissive, balanced and strict) were investigated. The permissive type uses either the combination of sender anomaly or drift with at least one other supporting suspicious cue to preserve high recall with minimal increase in precision. Using this configuration, isolation forest achieves a precision of 0.26, a recall of 0.72 and an F-1 score of 0.38. Additionally, One Class SVM achieves a precision of 0.26, a recall of 0.69 and an F-1 score of 0.38.

A second type of warning (balanced) requires two suspicious cues prior to generating a warning. Both models experience a significant decrease in their recall value to around 0.11 and a corresponding decrease in their F-1 scores to around 0.15. This clearly shows that most synthetic anomalous emails exhibit only one primary suspicious cue at a given time and thus would cause a two cue threshold to be unfeasible for use in this proof-of-concept.

The third type of warning (strict), requiring both anomaly and drift to have occurred with at least one supporting suspicious cue generates the highest precision for all the warnings analyzed. Using this configuration, isolation forest achieves a precision of 0.28, a recall of 0.49 and an F-1 score of 0.35. While this configuration causes a loss in recall compared to the previous configurations, it represents the largest gain in warning selectivity and consequently matches the goal of providing a simple sender style based warning most closely.

As shown in Figure 2, the permissive warning mode preserved higher recall, whereas the strict mode produced the strongest precision albeit, not by much.

C. Summary of Experiments Findings

Together, the experimental findings demonstrate a distinct difference in performance between the labeled stylometric classification and sender specific anomaly detection tasks. As shown in Experiment 1, the feature set performs exceptionally well when both legitimate and synthetic anomalous emails are available during training. Conversely, Experiment 2 demonstrates that sender-relative anomaly detection is considerably more difficult when synthetic anomalous emails are designed to resemble a senders legitimate style as closely as possible.

The anomaly detection configurations do not match the performance levels obtained by the supervised learners. However, the warning rule analysis demonstrated that sender relative style deviations may provide some discriminative information when used in conjunction with supporting suspicious cues. Of the anomaly detection configurations examined, the strict warning configuration resulted in the best precision and worst recall while the permissive configuration had the worst precision and best recall.

V. Conclusion

In this research we explored if sender-style communication drift can be used to differentiate suspicious e-mail communications from legitimate organizational communications. We created synthetic anomalous e-mails based on e-mail data derived from Enron E-mail Data Set using a gpt's 4o-mini language model. We then tested both stylometric and sender-relative-drift feature evaluations to determine how well they distinguished between legitimate corporate communications and the synthetic anomalous communications. Our supervised classification test found that our sender-conditioned synthetic anomalous communications features held strong discriminative power (and therefore high predictive accuracy) within a labeled context. Furthermore, the Random Forest model performed the best across all models in terms of overall performance. The most interesting discovery was that distinguishing suspicious communications through the sole use of sender-specific writing style differences is clearly a harder task than distinguishing

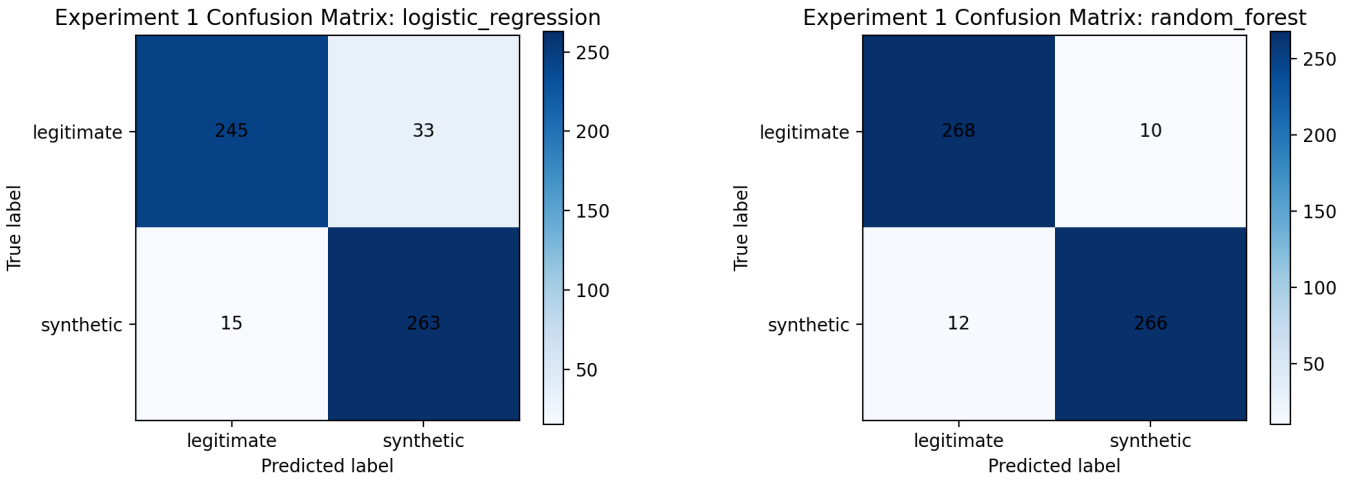

Fig. 1. Confusion matrices for Experiment 1 supervised classification. Logistic Regression produced more false positives and false negatives than Random Forest, while Random Forest achieved the strongest overall classification performance.

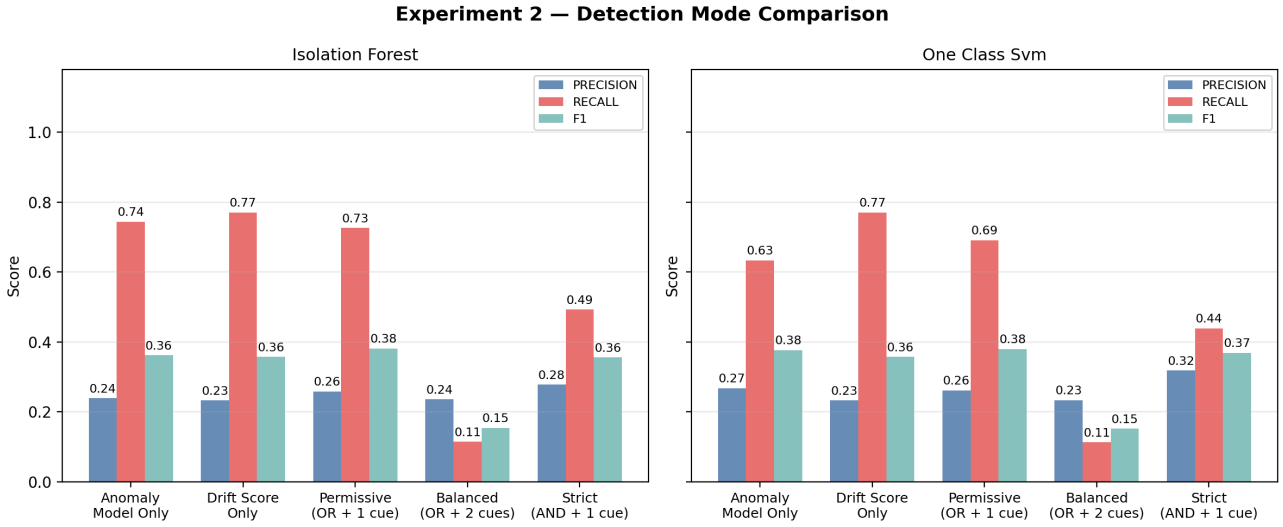

Fig. 2. Comparison of Experiment 2 detection modes across Isolation Forest and One-Class SVM. Permissive warning rules preserved higher recall, while strict warning rules improved precision at the cost of reduced recall. Balanced mode was overly restrictive and substantially reduced recall for both models.

those same communications through the presence of anomalous stylistic patterns.

Our findings support that sender-style communication drift is indeed a meaningful signal; however, it is still not defined enough to serve as a single mechanism for detecting phishing attacks in this proof-of-concept. A more useful application would appear to be as an added-layer warning mechanism that identifies potentially anomalous communications by determining whether the communication differs significantly in its writing style compared to what the sending party has previously written. Furthermore, a combination of this method with other types of suspiciousness indicators (such as time-sensitive requests for action regarding financial matters) could enhance the effectiveness of lightweight sender-based warnings that are designed to alert users of potential phishing attempts without substituting for existing anti-phishing mechanisms employed by email providers.

We also have many limitations in our study. Rather than collecting anomalous e-mails from actual phishing events, we synthesized them ourselves. Additionally, because of some issues with style alignment filtering,

we ultimately ended up with fewer total examples in the final synthetic dataset than had been initially anticipated. Finally, we constructed a ”Drift Score” for the anomaly detection testing portion of this study as a simple, intuitive example of a way to measure how far an email's writing deviates relative to its writer. While we believe this research represents a reasonable starting point for measuring such deviations, there are certainly better ways to do so that will likely emerge during future studies.

Future studies should examine richer representations of stylometry, create more precise warning thresholds for sender-based warning systems, and develop larger collections of real-world phishing e-mail communications.

REFERENCES

[1] Softonic, “Phishing Email Security: Stats and Trends,” Softonic, 2024. [Online]. Available: https://en.softonic.com/business/article/phishing-email-security-stats-trends

[2] C. Opara, P. Modesti, and L. Golightly, “Evaluating Spam Filters and Stylometric Detection of AI-Generated Phishing Emails,” Expert Systems with Applications, Jun. 2025.

[3] M. Popa, “Authentication of Electronic Messages Using Stylometry,” U.S. Patent 11,516,223 B2, Dec. 20, 2022. [Online]. Available: https://patents.google.com/patent/US11516223B2/en

[4] “A Machine Learning Approach for Phishing Detection,” Expert Systems with Applications, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417425006669

[5] P. Dewan, A. Kashyap, and P. Kumaraguru, “Analyzing Social and Stylometric Features to Identify Spear Phishing Emails,” arXiv preprint arXiv:1406.3692, Jun. 2014.

[6] A. Kao, “LLM Spear Phishing,” 2023. [Online]. Available: https://andrew-kao.github.io/files/llmspearphish.pdf

[7] “Phishing Detection Using Advanced AI Techniques,” Expert Systems with Applications, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417426004203