ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 2. С. 417–423
DOI: https://doi.org/10.5281/zenodo.24311937
Сайт журнала: https://cajournal.kz

УДК 004.056

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА СИСТЕМЫ ОБНАРУЖЕНИЯ AI-GENERATED PHISHING-АТАК НА ОСНОВЕ ПОЛНОГО ФАКТОРНОГО ПЛАНА 2^3: ИМИТАЦИОННОЕ ИССЛЕДОВАНИЕ

Н.Т. Нурахунова, А.С. Канетей
Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Институт автоматики и информационных технологий, кафедра кибербезопасности, Алматы, Казахстан

Нурахунова Наиля Тайиржановна — студент 4 курса, Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Институт автоматики и информационных технологий, кафедра кибербезопасности, Алматы, Казахстан.

Канетей Аида Сапаркызы — студент 4 курса, Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Институт автоматики и информационных технологий, кафедра кибербезопасности, Алматы, Казахстан.

Научный руководитель: Абдуллаева Асель Сейдуллаевна — Казахский национальный исследовательский технический университет имени К.И. Сатпаева.

© Н.Т. Нурахунова, А.С. Канетей, 2026

Аннотация: Генеративные языковые модели позволяют создавать грамматически корректные и персонализированные фишинговые письма, что снижает эффективность детекторов, опирающихся на орфографические ошибки или известные шаблоны. Цель работы – показать планирование и статистическую обработку многофакторного эксперимента для оценки системы обнаружения фишинга; построен полный факторный план 2^3 с тремя факторами: NLP-признаки (A), стилометрические и AI-origin признаки (B) и новый генератор текста, не представленный при обучении (C); отклик – F1. Поскольку реальный классификатор не запускался, значения F1 для 24 наблюдений (восемь условий × три повтора) имитационные и служат только для демонстрации методики. Регрессионная модель, полученная методом наименьших квадратов ($R^2 = 0{,}988$; $F(7,16) = 190{,}6$; $p < 0{,}05$), показала положительное влияние факторов A (+0,020) и B (+0,031) и отрицательное влияние фактора C (−0,025); план ортогонален (VIF = 1), для трёхшагового прогноза использован линейный тренд. Результаты иллюстрируют методику и должны быть заменены измерениями на реальном размеченном корпусе.

Ключевые слова: обнаружение фишинга, генеративный искусственный интеллект, NLP, стилометрия, полный факторный эксперимент, регрессия, F1, прогнозирование

1. Введение

Фишинг остаётся одним из наиболее распространённых начальных векторов киберинцидентов, поскольку он нацелен на человеческий фактор, а не на уязвимость программного обеспечения. Распространение больших языковых моделей (LLM) изменило экономику таких атак: грамматически корректные, стилистически выдержанные и персонализированные сообщения могут быть созданы за секунды и на любом языке [1]. Обнаружение, основанное на орфографических ошибках, неуклюжих формулировках или известных шаблонах, поэтому больше не является достаточным.

Более перспективным направлением является объединение нескольких источников свидетельств: технических атрибутов сообщения (заголовки, URL, результаты аутентификации), семантического анализа его содержимого с помощью обработки естественного языка (NLP), стилометрических и AI-origin индикаторов, а также контекста коммуникации. Наше более раннее концептуальное исследование [2] утверждало, что такое объединение необходимо и что обнаружение авторства ИИ не должно заменять обнаружение фишингового намерения, поскольку легитимные отправители также используют ИИ-ассистенты. Остаётся количественно оценить вклад каждого блока признаков и его надёжность, когда злоумышленники переключаются на генератор текста, который детектор не видел во время обучения.

Планирование многофакторного эксперимента является подходящим инструментом для этой задачи: оно оценивает главные эффекты и взаимодействия с небольшим числом запусков [3, 4]. Следует сразу подчеркнуть, что описываемый здесь эксперимент является имитационным: значения F1 представляют собой демонстрационные данные, а не выход обученного классификатора.

1.1. Цель и задачи

Цель состоит в том, чтобы на имитационном наборе данных продемонстрировать планирование и статистическую обработку эксперимента, оценивающего влияние NLP- и стилометрических признаков на качество обнаружения фишинга при изменении генератора текста. Задачи заключаются в определении факторов и отклика; построении матрицы плана 2^3; подборе регрессионной модели методом наименьших квадратов; построении поверхностей отклика и линий уровня; проверке корреляции и мультиколлинеарности; подборе линейного тренда и выполнении трёхшагового прогноза; а также в оценке адекватности модели и указании ограничений.

2. Материалы и методы

2.1. Объект исследования и отклик

Объектом исследования является концептуальная система обнаружения фишинговых электронных писем, в том числе написанных с помощью генеративного искусственного интеллекта. Откликом является F1-мера – гармоническое среднее точности (Precision, P) и полноты (Recall, R):

$$F1 = \frac{2PR}{P+R} \quad (1)$$

которая лежит в диапазоне от 0 до 1; более высокое значение указывает на лучший баланс между точностью положительных предсказаний и полнотой обнаружения. Значения F1 смоделированы так, чтобы напоминать правдоподобный сценарий, и предназначены для отработки расчётов. В реальном исследовании они должны быть заменены результатами запуска классификатора на размеченном корпусе.

2.2. Факторы и план

Использованы три двухуровневых фактора (Таблица 1). Фактор A добавляет NLP-анализ текста к базовым техническим признакам (заголовки и URL); фактор B добавляет стилометрический и AI-origin блок; фактор C описывает, присутствовал ли тестовый генератор в обучающих данных.

Таблица 1 – Факторы и уровни

Фактор	Уровень −1	Уровень +1
A – NLP-признаки	Только технические признаки (заголовки, URL)	Технические признаки + NLP-анализ текста
B – стилометрия / AI-origin	Без стилометрического блока	Добавлены стилометрические и AI-origin признаки
C – тестовый генератор	Генератор представлен в обучающей выборке	Новый генератор не представлен при обучении

При трёх факторах на двух уровнях полный факторный план имеет N = 2^3 = 8 условий. Три имитационных повтора для каждого условия дают 24 наблюдения (Таблица 2). Кодированные уровни (−1, +1) делают эффекты сопоставимыми, а матрицу плана – ортогональной.

2.3. Регрессионная модель и метод наименьших квадратов

Полная модель содержит главные эффекты и все взаимодействия:

$$F1 = b_0 + b_1 A + b_2 B + b_3 C + b_{12} AB + b_{13} AC + b_{23} BC + b_{123} ABC \quad (2)$$

Коэффициенты оценивались методом наименьших квадратов, то есть минимизацией суммы квадратов остатков:

$$SSE = \sum_{i=1}^{24} \left(F1_i - \hat{F}1_i \right)^2 \quad (3)$$

Значимость отдельных членов проверялась с помощью t-статистики (16 степеней свободы остатков), а общая значимость – с помощью F-критерия Фишера. Адекватность характеризовалась коэффициентами R^2, скорректированным R^2, RMSE и графиком остатков.

2.4. Тренд и прогноз

Для демонстрации прогнозирования использовался отдельный имитационный ряд F1 за восемь последовательных итераций настройки: 0,720, 0,735, 0,748, 0,759, 0,771, 0,781, 0,792 и 0,801. Линейный тренд был подобран методом наименьших квадратов и экстраполирован на три шага вперёд. Ошибка подгонки оценивалась по RMSE на исходных восьми точках. Этот ряд является иллюстративным и не представляет реальных измерений.

3. Результаты и обсуждение

3.1. Имитационные результаты и эффекты факторов

Матрица плана и имитационные значения F1 приведены в Таблице 2.

Таблица 2 – Матрица плана и имитационные значения F1

№	A	B	C	F1 (п1)	F1 (п2)	F1 (п3)	Среднее
1	−1	−1	−1	0,747	0,759	0,761	0,756
2	+1	−1	−1	0,782	0,794	0,780	0,785
3	−1	+1	−1	0,807	0,812	0,799	0,806
4	+1	+1	−1	0,848	0,846	0,841	0,845
5	−1	−1	+1	0,695	0,683	0,698	0,692
6	+1	−1	+1	0,737	0,733	0,741	0,737
7	−1	+1	+1	0,755	0,764	0,757	0,759
8	+1	+1	+1	0,812	0,804	0,808	0,808

Наибольшее среднее значение F1 (0,845) наблюдается, когда включены оба блока (NLP и стилометрия), а генератор известен (A = +1, B = +1, C = −1). Переход к новому генератору снижает F1 во всех восьми условиях, иллюстрируя проблему переносимости. Примечательно, что комбинированный набор признаков ограничивает это ухудшение: падение от C = −1 к C = +1 составляет 0,064 для базового варианта (A = B = −1), но только 0,037, когда присутствуют оба блока (A = B = +1). Эквивалентно, прирост от добавления NLP-признаков составляет 0,034 на известном генераторе и 0,047 на новом, а прирост от стилометрии – 0,055 и 0,069 соответственно.

Подгонка методом наименьших квадратов дала уравнение:

$$\hat{F}1 = 0{,}7735 + 0{,}0204\,A + 0{,}0310\,B - 0{,}0245\,C + 0{,}0017\,AB + 0{,}0032\,AC + 0{,}0035\,BC - 0{,}0006\,ABC \quad (4)$$

Коэффициенты A и B положительны: в имитационных данных добавление NLP- и стилометрических признаков повышает F1; отрицательный коэффициент C отражает потери на невидимом генераторе. Поскольку факторы закодированы как ±1, полный эффект фактора (изменение между его уровнями) равен двум коэффициентам: около 0,041 для A, 0,062 для B и −0,049 для C. Члены взаимодействий малы по сравнению с главными эффектами; AC и BC значимы при α = 0,05, тогда как AB и ABC – нет (Таблица 3). Значимые члены AC и BC

согласуются с приведённым выше наблюдением о том, что оба блока признаков помогают сильнее на новом генераторе.

Таблица 3 – Регрессионные коэффициенты (общая стандартная ошибка ≈ 0,0012)

Член	Коэффициент	t (16 ст. св.)	p
b_0	0,7735	631,5	<0,001
A	0,0204	16,64	<0,001
B	0,0310	25,28	<0,001
C	−0,0245	−20,04	<0,001
AB	0,0017	1,39	0,182
AC	0,0032	2,62	0,019
BC	0,0035	2,82	0,012
ABC	−0,0006	−0,51	0,617

3.2. Поверхность отклика и линии уровня

Поверхности (Рисунок 1) и графики линий уровня (Рисунок 2) построены на основе подобранной модели. Для известного генератора комбинация обоих блоков даёт максимальный F1. Для нового генератора поверхность располагается ниже для любой комбинации A и B, но добавление NLP и стилометрии по-прежнему вносит положительный вклад. Поверхности близки к параллельным плоскостям, что согласуется с малыми коэффициентами взаимодействия.

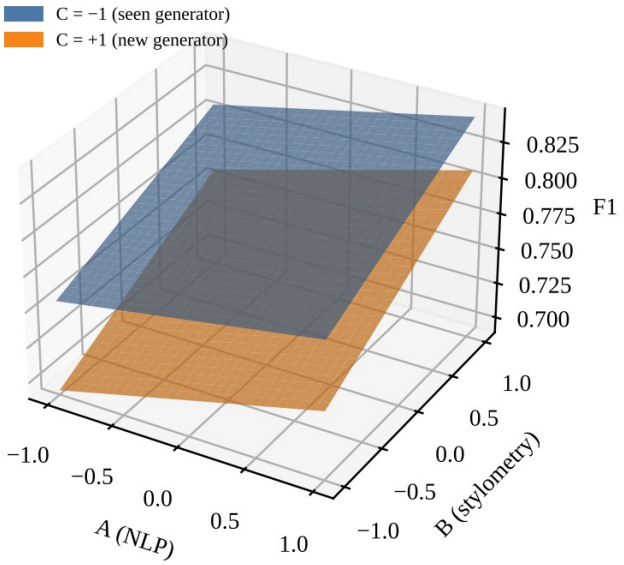


Рисунок 1 - Поверхность отклика F1 для C = −1 и C = +1

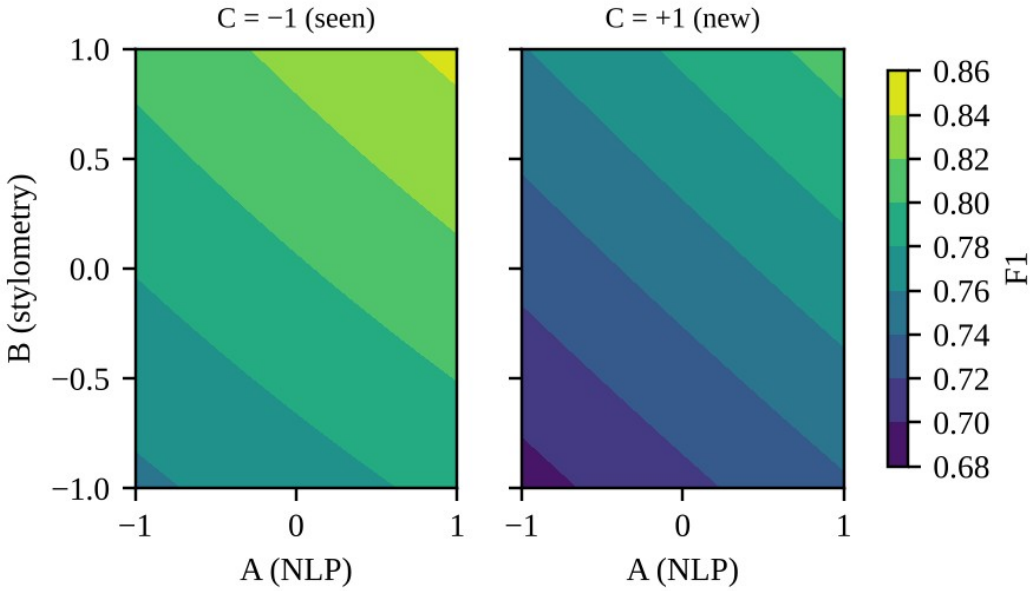


Рисунок 2 - Графики линий уровня F1 для известного (слева) и нового (справа) генератора

3.3. Корреляция и мультиколлинеарность

Для кодированной матрицы полного факторного плана корреляции между главными факторами равны нулю: r(A,B) = r(A,C) = r(B,C) = 0. Все семь столбцов модели (A, B, C, AB, AC, BC, ABC) взаимно ортогональны, поэтому оценки коэффициентов не дублируют друг друга, а все коэффициенты инфляции дисперсии (VIF) равны 1; в матрице плана нет мультиколлинеарности. Важно отличать корреляцию между факторами от связи между факторами и откликом: нулевая корреляция между A, B и C не означает, что они не влияют на F1; здесь главные эффекты велики и имеют чёткие знаки.

3.4. Тренд и трёхшаговый прогноз

Линейный тренд имитационного ряда настройки имеет вид (t – номер итерации):

$$\hat{F}1 = 0{,}7118 + 0{,}0115\,t \qquad (5)$$

Прогнозы для итераций 9, 10 и 11 составляют 0,815, 0,826 и 0,838 (Рисунок 3). RMSE подгонки по восьми исходным точкам составляет 0,0018. Поскольку ряд короткий и имитационный, прогноз лишь демонстрирует метод; практическое использование требует реальных последовательных измерений и валидации на отложенных данных. Линейная экстраполяция также не может продолжаться бесконечно, так как F1 ограничен единицей, а приросты обычно насыщаются.

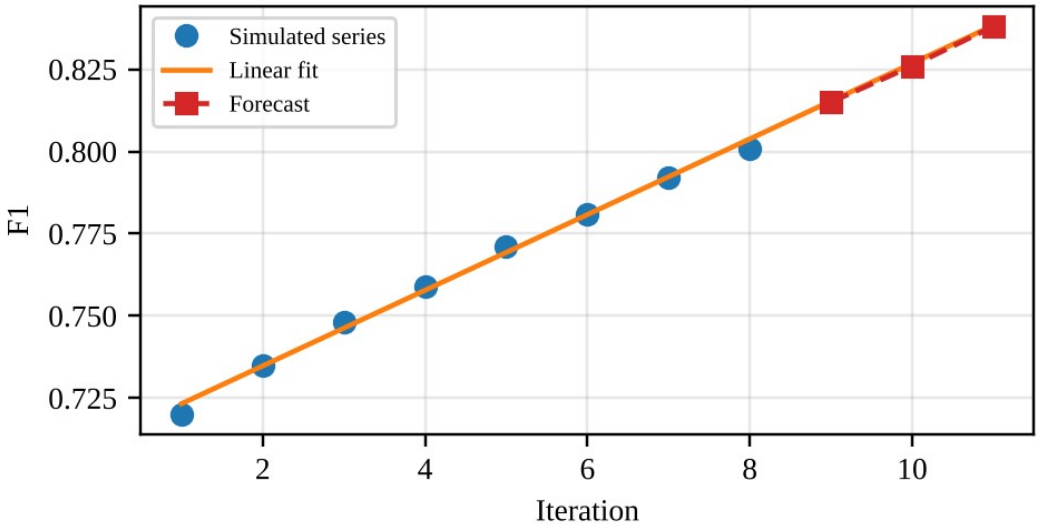


Рисунок 3 - Линейный тренд и трёхшаговый прогноз F1

3.5. Адекватность модели

Показатели качества модели приведены в Таблице 4.

Таблица 4 – Показатели качества модели

Показатель	Значение
Наблюдений	24 (8 условий × 3 повтора)
R^2	0,9882
Скорректированный R^2	0,9830
$F(7,16)$	190,61
p (критерий Фишера)	$3{,}36 \times 10^{-14}$
RMSE	0,0049

F-критерий Фишера отвергает при α = 0,05 гипотезу о том, что все регрессионные коэффициенты равны нулю ($F(7,16) = 190{,}61$, $p = 3{,}36 \times 10^{-14}$). $R^2 = 0{,}988$ показывает, что модель очень хорошо описывает построенные данные; это не подразумевает аналогичной точности на реальных электронных письмах. Остатки (Рисунок 4) малы, лежат по обе стороны от нуля и не показывают явной кривизны, но при 24 имитационных наблюдениях визуальная проверка не может заменить полную диагностику на реальном наборе данных.

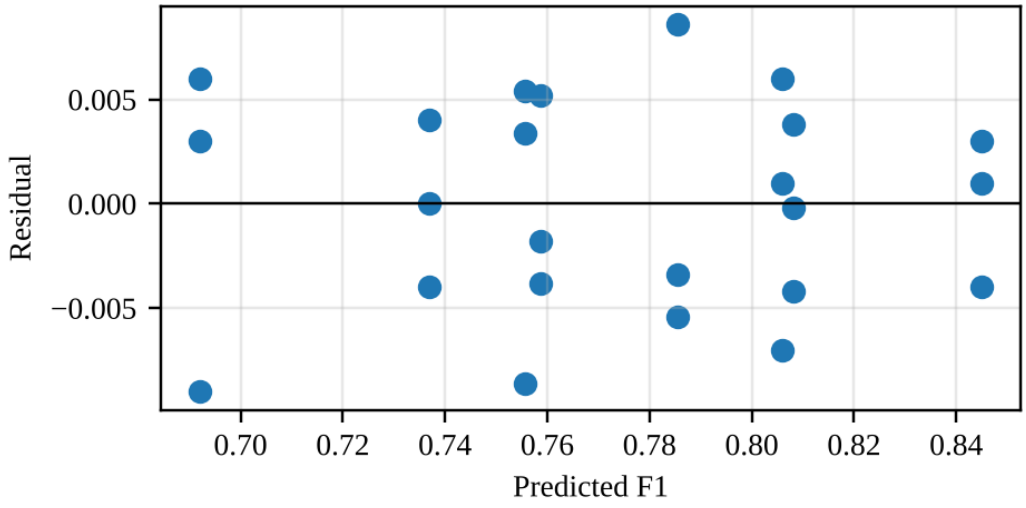


Рисунок 4 - Остатки в зависимости от предсказанного F1

3.6. Ограничения и дизайн реального эксперимента

Главным ограничением является имитационный характер данных: классификатор не запускался на реальном корпусе и фактическая разметка не проводилась. Поэтому коэффициенты, R^2 и прогноз не могут использоваться в качестве доказательства эффективности готовой системы. В реальном эксперименте следует использовать четыре класса сообщений: легитимные, написанные человеком; легитимные с поддержкой ИИ; фишинговые, написанные человеком; и фишинговые, сгенерированные ИИ. Данные должны быть разделены по времени и генератору, дубликаты удалены, а версии моделей зафиксированы. Помимо F1, должны сообщаться Precision, Recall, PR-AUC [5] и уровень ложноположительных срабатываний, поскольку детектор, блокирующий легитимную почту с поддержкой ИИ, на практике неприемлем. Вывод о том, что обнаружение происхождения ИИ должно дополнять, а не заменять обнаружение фишингового намерения, согласуется с нашим более ранним концептуальным исследованием [2].

4. Заключение

Был построен полный факторный план 2^3 для исследования системы обнаружения AI-generated phishing-атак с тремя факторами (NLP-признаки, стилометрические признаки, новый тестовый генератор) и откликом F1. Регрессионная модель с главными эффектами и взаимодействиями была подобрана методом наименьших квадратов по 24 имитационным наблюдениям ($R^2 = 0{,}988$; $F = 190{,}61$; $p < 0{,}05$); кодированный план ортогонален с VIF = 1.

В имитационных данных оба блока признаков увеличивают F1 (коэффициенты +0,020 и +0,031), невидимый генератор уменьшает его (−0,025), а комбинированный набор признаков снижает потери, вызванные сменой генератора, с 0,064 до 0,037. Линейный тренд отдельного имитационного ряда дал трёхшаговый прогноз 0,815, 0,826 и 0,838.

Основной результат носит методологический характер: продемонстрирован полный цикл планирования, обработки и оценки факторного эксперимента. Практическое следствие заключается в обосновании необходимости тестирования дополнительных NLP- и стилометрических признаков всегда на невидимых генераторах. Подтверждение требует эксперимента на реальном, размеченном и репрезентативном корпусе электронных писем.

Список литературы:

1. Heiding F., Schneier B., Vishwanath A., Bernstein J., Park P.S. Devising and Detecting Phishing Emails Using Large Language Models // IEEE Access. 2024. Vol. 12. P. 42131–42146.

2. Nurakhunova N.T., Kanetei A.S. AI-generated phishing attack detection system [Laboratory work No. 2, Fundamentals of student research]. Almaty: Satbayev University, 2026.

3. Box G.E.P., Hunter J.S., Hunter W.G. Statistics for Experimenters: Design, Innovation, and Discovery. 2nd ed. Hoboken: Wiley, 2005.

4. Montgomery D.C. Design and Analysis of Experiments. 9th ed. Hoboken: Wiley, 2017.

5. Saito T., Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets // PLOS ONE. 2015. Vol. 10, No. 3. Art. e0118432.