

Sift: AI-Assisted SMS Scam Detector

Can Jin

Gwen Guan

10/06/2026

1. Executive Summary

Sift is an AI-assisted tool that helps international students check whether a text message may be a scam. These students may not be familiar with U.S. institutions, typical communication practices, or common types of scams. A user can paste an English SMS into the interface to see a Scam or Normal prediction, the model score, any warning signs that were found, and an optional explanation in plain language.

The starting dataset contained 500 labeled messages: 250 Scam messages and 250 Normal messages. The data was split before augmentation into 350 training, 75 validation, and 75 test messages. An additional 500 messages were generated only from the training set, producing a final training set of 850 messages while keeping the validation and test sets original and untouched.

Five TF-IDF and Logistic Regression configurations were compared using validation data. Rule-based checks also look for suspicious links, urgent wording, payment requests, and threats. If enabled, Qwen uses these structured results to produce a plain-language explanation.

Sift is meant to help users make their own judgment, not make the decision for them. Its main limitations include its small English-only dataset, changing scam patterns, and inability to verify the actual sender.

2. Introduction and Motivation

International students often have to judge messages from unfamiliar U.S. institutions. A text may appear to come from a bank, university, delivery company, embassy, or immigration office. Differences in language, local procedures, and communication styles can make it harder to decide whether a request is normal. Urgent wording or threats can add pressure and lead a student to respond before checking the sender. [1]-[4]

When a suspicious message arrives, a student may search the phone number, inspect the link, ask someone else, or contact the organization directly. This process can be slow and inconsistent, especially when the message seems urgent. A wrong decision can have two different consequences: trusting a scam may expose money or personal information, while rejecting a real message may cause the student to miss an important request. [1]-[4]

Therefore, Sift was created as a decision-support tool for this purpose. The user pastes an English SMS into the interface. The system returns a Scam or Normal prediction, a model score, and warning signs found in the message. An optional language model then explains the structured result in plain language. The system does not verify the sender, block messages, or make the final decision.

The main technical goal was to recognize scam messages while limiting unnecessary warnings. Performance was measured with accuracy, precision, recall and a confusion matrix. We defined success in three ways: high classification accuracy on the untouched test set, a clean and easy-to-follow interface, and results that help users make a judgment by showing both the prediction and the reasons behind it. The final decision remains with the user.

3. Design Process

3.1 Data Ethics and Responsible Use

The dataset focuses on English messages and scams relevant to international students, so it may miss other languages, locations, student groups, and new scam patterns. Source labels may also contain errors. Sift cannot verify the sender or prove that a message is fraudulent. Its score provides supporting information, while the user makes the final decision.

No separate redacted copy was created. Some records retain names, phone numbers, email addresses, organization names, or live-looking URLs because these details can help the classifier recognize message patterns. The system treats them only as text and does not open or contact them. This limits safe data sharing, so future work should replace such details with consistent placeholders and compare model performance before and after redaction.

3.2 Design Priorities and System Structure

Our first priority was to identify scam messages without missing harmful examples. The second priority was to limit false warnings because too many incorrect alerts could reduce user trust. The third priority was to keep the system fast, understandable, and easy to inspect.

We separate classification, warning-sign detection, and explanation into different components. TF-IDF and Logistic Regression produce the Scam or Normal prediction and model score. Rule-based checks find suspicious wording, links, payment requests, and threats, but they do not change the classification. Qwen is used only to explain the structured result in plain language. This separation keeps the prediction consistent and prevents the language model from silently changing the classifier's decision.

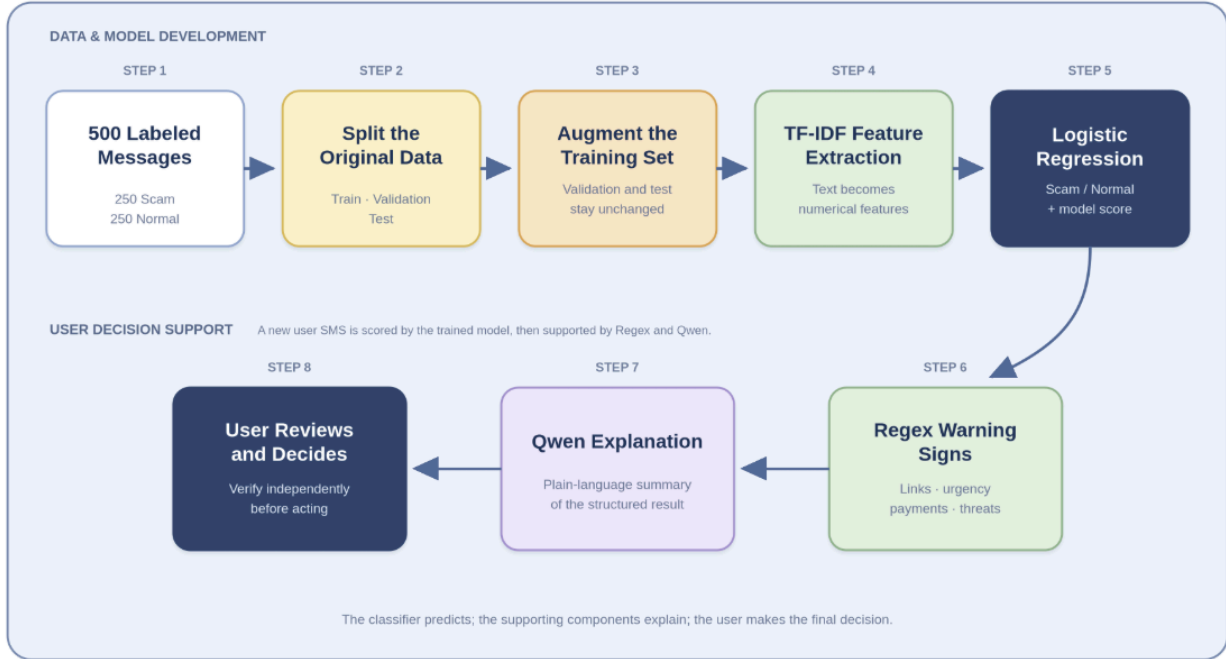


Figure 1. Sift development and user workflow. The upper path shows data preparation and model training. The lower path shows how a user’s message moves through classification, warning-sign detection, explanation, and final review. [5]

This structure involves a tradeoff. A larger transformer classifier could capture more context, but it would require more training data, computing time, and resources. TF-IDF and Logistic Regression were selected because they are faster, easier to inspect, and appropriate for the size of this dataset.

3.3 Data Collection and Labeling

The base dataset contained 500 labeled message examples: 250 labeled Scam and 250 labeled Normal. The 250 Scam messages included 150 records selected from Mishra and Soni’s SMS Phishing Dataset [6], 92 records from the NCSU WSPR SMS Phishing Dataset [7], four examples from the University of Iowa [1], two from UC Davis [2], one from RPI [3], and one from UC Berkeley [4].

For the Normal dataset, we collected examples based on official information from banks, telecommunications companies, and other institutions, including ANZ [10], ITV [11], U.S. Bank [12], O2 [13], and Unity Bank [14]. These sources were used to create or adapt Normal messages. We also included 12 ham messages from a publicly available dataset [15]. The original Scam and Normal datasets, including their source information and URLs where available, are provided in Appendices A.1 and A.2, respectively.

The Mendeley dataset is published under the CC BY 4.0 license [6]. The NCSU WSPR repository uses the MIT License [7].

The selected Scam messages covered situations that international students may encounter, including banking, universities, delivery services, government impersonation, immigration threats, gift-card requests, and family impersonation. Label 0 represents a routine Normal message.

Each record received a unique ID, message text, and binary label:

Label	Class	Definition
0	Normal	A routine message
1	Scam	A message that attempts to obtain money, credentials, personal information, or an unsafe response through deception or impersonation

Messages were reviewed for English-language content and relevance to the project. Blank messages, extremely short messages, exact normalized duplicates, and records with conflicting labels were removed. Source information was preserved when available so that records could be traced back to their original datasets or webpages.

Appendix A contains the complete 250 Scam and 250 Normal base records used by the project.

3.4 Data Splitting and Leakage Prevention

The 500 base messages were divided before any augmentation took place. This prevented generated versions of a message from appearing in validation or test data.

Dataset stage	Scam	Normal	Total
Original training set	175	175	350
Validation set	37	38	75
Test set	38	37	75
Augmented training messages	250	250	500
Final training set	425	425	850

Only the 350 original training messages were allowed to serve as augmentation parents. The validation and test sets remained original and untouched. Record IDs and parent IDs were checked to confirm that no validation or test message was used to generate training data. Exact normalized duplicates and conflicting labels were also checked after the final datasets were created.

This split allowed the validation set to guide model and threshold selection while reserving the test set for one final evaluation.

Appendix B provides the augmentation records, including the parent IDs, labels, and augmentation methods.

3.5 Training-Set Augmentation

Five hundred additional messages were created from the original training set to increase wording variation while preserving the original labels. The augmentation process produced 200 contextual paraphrases, 120 entity-substitution examples, 100 controlled typing variations, and 80 SMS-style variations.

Contextual paraphrasing used FLAN-T5-small [8] to rewrite a message while keeping its original meaning and label. Entity substitution changed details such as organization names, amounts, dates, or locations. Controlled typing variation introduced limited spelling and formatting changes. SMS-style variation shortened or reformatted messages to resemble common texting patterns.

Generated messages were rejected if they were blank, too short, too long compared with their parent message, or exact duplicates of existing records. Protected values were preserved when necessary.

All 350 original training messages were used as augmentation parents: 200 produced one new message, while 150 produced two. Spreading the 500 new examples across the full training set reduced dependence on a small group of parent messages.

Appendix B also records the parent ID and augmentation method for each generated message.

3.6 Model Training and Selection

TF-IDF was used to convert each SMS into numerical features. Word features represented individual words and two-word phrases. Character features represented short character patterns that can help the model recognize spelling variations and obfuscated words.

Logistic Regression used these features to estimate a score for the Scam class. The score was then compared with a classification threshold to produce the final Scam or Normal prediction.

Five configurations were trained and compared. The first configuration used word unigrams only, and the second used word unigrams and bigrams. The remaining three configurations combined word and character features while testing Logistic Regression values of $C=0.5$, $C=1.0$, and $C=2.0$.

The configurations were ranked by validation average precision, followed by validation F1, validation recall, and training time. The three word-and-character configurations produced the same values for the first three selection metrics. In the saved run, word_char_c2 (configuration

combined word-level and character-level TF-IDF features with Logistic Regression using $C=2.0$) had the shortest training time among those tied configurations and was therefore selected.

After the model configuration was fixed, thresholds from 0.10 to 0.90 were evaluated using the validation set. Thresholds of 0.38 and 0.39 produced the same F1, recall, and precision, so 0.38 was selected as the lower value. The test set remained unused until both the model configuration and threshold had been fixed.

3.7 Model Evaluation

After the configuration and threshold were fixed, the selected model was evaluated once on the untouched test set.

Metric	Validation Set	Test Set
Accuracy	97.3%	98.7%
Precision	97.3%	97.4%
Recall	97.3%	100%

The test confusion matrix contained 36 true negatives, 38 true positives, one false positive, and no false negatives.

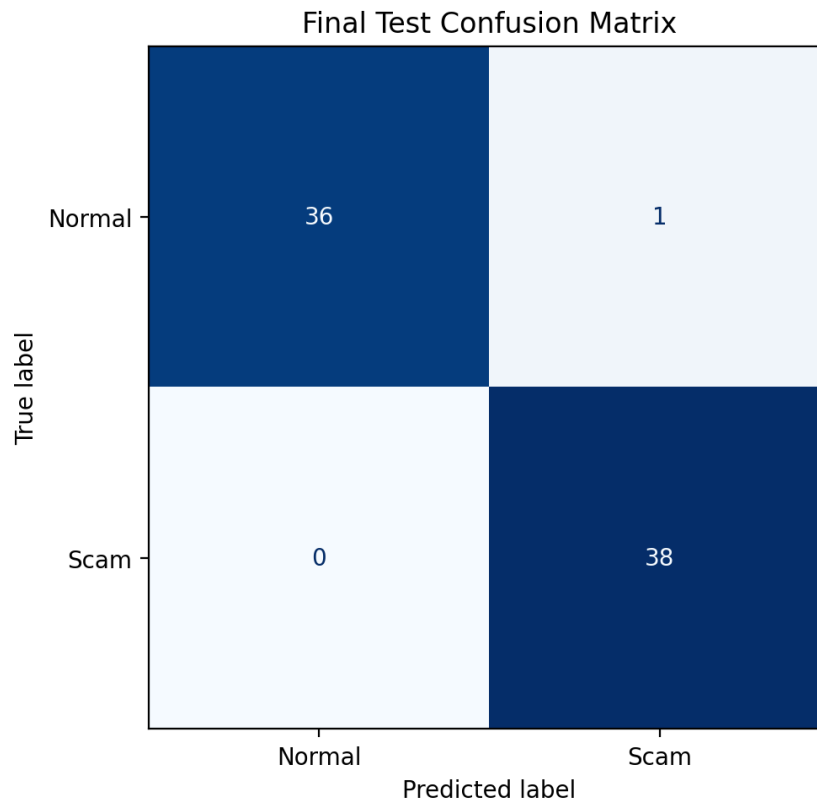


Figure 2. Confusion matrix for the final model on the untouched 75-message test set.

The only false positive was: “Your package has been shipped. Track your delivery at this link.”

The message contains delivery and link-related wording that also appears in scam messages, so the model assigned it a score above the 0.38 threshold. This example shows the main tradeoff created by the lower threshold: the model detected every scam message in this test set, but one legitimate message was flagged.

3.8 Rule-Based Warning-Sign Detection

We used regular expressions to identify common warning signs in SMS messages. The `detect_signals()` function checks five main categories: payment requests, password or verification code requests, urgent or threatening wording, requests to follow links, and references to institutions or family members. For each match, the system records the category, the matched text, and its position in the message. The interface then highlights these words so users can see what triggered the rule.

We also added checks to reduce incorrect warnings. For example, "Never share your verification code" is a protective reminder, not a request for personal information. Similarly, mentioning a bank or university does not necessarily mean the message is a scam. The system separates these matches into warning signs, contextual information, and protective wording.

For messages containing URLs, the system checks whether a link uses a shortened URL, an IP address, a username in the URL, or an unusual domain structure. These patterns may deserve closer review, but they do not prove that a website is unsafe. The system only examines the link text and does not open or verify the website.

Regex provides supporting text evidence and does not change the Logistic Regression prediction or score. This separation keeps the classification consistent while giving users more information about the message. But, fixed rules may miss unfamiliar wording or misunderstand context, so the detected signs should not be treated as proof of fraud.

3.9 Qwen Explanation and Output Validation

We used the off-the-shelf Qwen2.5-1.5B-Instruct model [9] to generate short explanations of SMS messages. We selected a relatively small instruction-following model to limit computing requirements and avoid training another model. Qwen was loaded using its pretrained weights without additional fine-tuning.

Qwen receives the cleaned SMS message, the fixed classifier result, the model score, the threshold, and some detected text cues. We designed the prompt to ask for one or two short English sentences, with a maximum of 50 words. The explanation should describe what the message says or asks the recipient to do, without inventing information.

One challenge with using Qwen is that it may make assumptions about risks that are not supported by the message. To reduce this problem, we improved the prompt and instructed Qwen not to invent threats, fees, deadlines, links, or requests for personal information. We also instructed it not to claim that a sender or website had been verified. The prompt asks Qwen not to treat normal wording, such as a bank name or promotional language, as evidence of fraud by itself. Qwen only provides explanations and cannot change the classifier's Scam or Normal prediction.

We added an output validation function, `check_explanation()`, to check generated explanations before displaying them. It checks whether quoted text appears in the original SMS, whether numbers and time units are supported by the message, and whether the explanation makes certain unsupported claims. It also checks that the response does not directly instruct users to pay, click a link, or share sensitive information.

If an explanation fails these checks, the system gives Qwen one more attempt using the failed checks as feedback. If the second attempt also fails, the interface displays "AI explanation unavailable" instead of presenting an unchecked explanation. The original **classification result** remains unchanged. Users can also turn off the AI explanation without affecting classification.

3.10 Interface and Interaction Design

We built Sift as a web application using Gradio. Our goal was to make the system simple enough for students to use without technical knowledge. Users can paste an English SMS message into the text box or choose one of four example messages, including a bank alert, gift-card request, campus reminder, and code protection message. They can also turn the Qwen explanation on or off before clicking "Analyze message."

The interface displays the Scam or Normal prediction, a model score between 0 and 1, and the decision threshold of 0.38. A visual bar shows whether the score is above or below the threshold. We also added a reminder that the score does not guarantee whether a message is safe. When the score is close to the threshold, the interface asks users to review the message more carefully.

The classification result appears before Qwen finishes generating its explanation, so users do not have to wait too long to see the result. Below the main result, the "Text evidence" section highlights words detected by regex rules. The "Model factors" section shows which TF-IDF features contributed most to the Logistic Regression score.

Users can also open the technical record to review the generated explanation and validation checks, or copy a short summary of the result. The interface encourages users to verify suspicious messages through official channels before taking action.

3.11 Error Handling, Safety, and System Boundaries

We added several checks to make the system more reliable when users enter messages or when a model fails. The `clean_message()` function removes extra spaces and checks the input before

classification. The system rejects empty or too-short messages with an error message. The Gradio input box limits messages to 2,400 characters, preventing users from entering longer messages.

For Qwen, we added output validation to reduce unsupported explanations. If the generated text fails the checks, the system retries once. If both attempts fail, the interface shows "AI explanation unavailable." We tested this fallback process using a synthetic test message, as shown in Appendix C.3. The classifier result is not changed by these failures, and users can still use Sift without enabling Qwen.

However, these checks cannot catch every problem. In one test, Qwen described an urgent callback request as a pressure tactic but also stated that there was no apparent risk. This unsupported safety judgment passed the automatic checks. The prompt also asks Qwen to keep explanations under 50 words, but the validator does not directly enforce this word limit. These limitations show that passing the checks does not guarantee a correct or consistent explanation.

We also considered privacy and safety. The interface reminds users to remove real names, account numbers, and verification codes before submitting messages. Messages are processed on the application server, and the application does not intentionally save them to a database. The system does not open suspicious links or verify sender identities. It also cannot guarantee that a message classified as Normal is safe.

Finally, Sift was developed using a relatively small English-language dataset, so its performance may be different for other languages or new scam patterns. Regex rules may also miss unfamiliar wording. For these reasons, we designed Sift as a tool to support user decisions rather than replace them.

4. User Experience and Workflow

4.1 User Workflow

Sift uses a single-page workflow. The user enters an English SMS, chooses whether to enable the AI explanation, and selects Analyze message. The system first checks the input and then displays a Scam or Normal prediction, the scam model score, and the saved classification cutoff. The user can also review the AI explanation and open the supporting information under Text evidence, Model factors, and How it works. After reviewing the result, the user decides whether the message should be trusted or verified through an independent official channel.

Figure 3 shows a completed analysis of a university study-group reminder. The model classified the message as Normal with a score of 0.3671, which is below the saved cutoff of 0.3800. Because the score is close to the cutoff, the interface displays an additional review cue instead of creating a different prediction class. The Qwen explanation summarizes why the message appears routine, allowing the user to see both the result and the reason behind it.

The screenshot displays the Sift app's analysis workflow. At the top, a header reads "Pause before you tap." with a subtitle "A closer look at the message. A clearer next step." Below this, the interface is divided into two main columns. The left column, titled "Take a closer look.", contains a text input field labeled "Message to analyze" with a placeholder "English messages. Remove real names, account numbers and codes first." The input field contains the text: "Reminder: your study group meets at the university library tomorrow at noon. Bring your notes." Below the input field, it shows "94 / 2,400 characters". Underneath is a section "Try an illustrative example" with four buttons: "Bank alert", "Gift card request", "Campus reminder", and "Code protection". At the bottom of this column is a checkbox labeled "Explain with AI" which is checked, with a note "Qwen runs on this app server. First use downloads the model." and a large dark blue button labeled "Analyze message →". The right column displays the results. At the top, it says "MODEL ASSESSMENT · SAVED CLASSIFIER". The prediction is "Normal" with a large score of "0.3671" and a subtitle "scam model score / 1". A horizontal progress bar shows the score is below the "Saved cutoff: 0.3800". A note states: "The score is close to the cutoff; this is a UI review cue, not a new class." Below this is a section titled "Qwen · AI explanation" with the text: "The message is a reminder about a study group meeting at the university library tomorrow at noon. It does not contain any apparent risks or scams."

Figure 3. Completed Normal-message analysis showing the original input, model prediction, scam score, classification cutoff, and Qwen explanation.

Figure 4 shows the interface after the result has been cleared. The status returns to Ready when you are, and the evidence panel states that no message has been analyzed. The same area provides access to supporting evidence, model factors, system information, and the technical record. A privacy notice also reminds users not to enter passwords, verification codes, or other sensitive information.

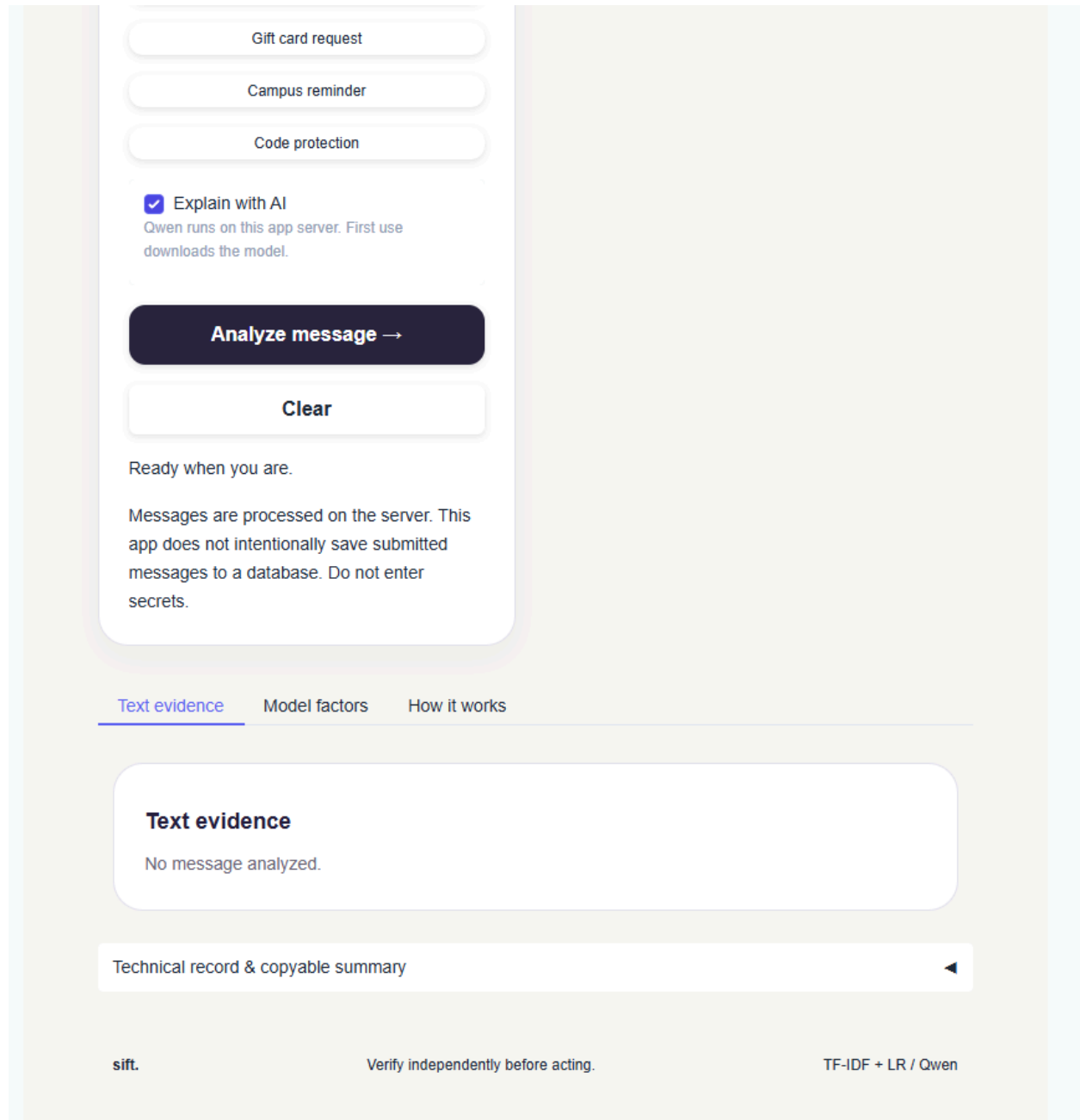


Figure 4. Cleared interface state showing the privacy notice, supporting-information tabs, and empty evidence panel.

Overall, Sift met our main classification goal, achieving 98.7% accuracy and 100% recall on the test set. The interface also provides a clear prediction and an optional AI explanation. However, Qwen can still produce unsupported explanations, and we have not conducted a formal usability study. More testing with real users is needed to evaluate how helpful and easy to understand the system is.

5. Reflection and Next Steps

We used generative AI throughout the project. It helped us draft most of the notebook code, debug errors, and organize the model workflow. During data preparation, it suggested ways to clean the messages, standardize labels, remove duplicates, and split the original data before augmentation to reduce data leakage. It also helped us understand why TF-IDF with binary Logistic Regression was a good choice for a small two-class text dataset: it trains quickly, works well with word and character patterns, and is easier to inspect than a large neural network.

AI worked best when the task had clear rules, such as checking duplicates, training model configurations, or calculating evaluation metrics. It was less useful for visual design. The first AI-generated versions of the Gradio interface were unattractive and poorly organized. It was difficult to explain the exact layout, spacing, colors, and visual hierarchy we wanted through text conversations, so the interface required several rounds of screenshots and revisions.

One issue we could not fully solve involved the Qwen explanation. In an actual test case, the classifier labeled a message as Scam, but Qwen said that the message showed no apparent risk. The current validator did not catch this conflict. Our next step is to make the validator check whether Qwen agrees with the classifier, uses the correct score and warning signs, and avoids unsupported claims.

We would also improve the quality and variety of the dataset by collecting more verified real messages and documenting their sources more carefully. The current original files were not redacted, so some records still contain names, phone numbers, email addresses, organization names, or live-looking URLs. We kept this wording because these details can help the classifier recognize message patterns, but it makes the dataset harder to share safely. We would also test the interface with more users and continue improving its layout.

6. References

- [1] University of Iowa Information Technology Services, “SMS Phishing Examples.” <https://its.uiowa.edu/services/phishing-and-anti-spam-efforts/sms-phishing-examples>
- [2] UC Davis Services for International Students and Scholars, “Common Scams.” <https://siss.ucdavis.edu/scams/common-scams>
- [3] Rensselaer Polytechnic Institute International Services for Students and Scholars, “Protect Yourself from Scams.” <https://isss.rpi.edu/resources/protect-yourself-scams>
- [4] UC Berkeley Information Security Office, “Phishing Examples Archive.” <https://security.berkeley.edu/education-awareness/phishing/phishing-examples-archive>
- [5] OpenAI, “ChatGPT.” Sift system workflow diagram generated with assistance from ChatGPT and edited by the authors, October 2026. <https://chatgpt.com/>
- [6] S. Mishra and D. Soni, “SMS Phishing Dataset for Machine Learning and Pattern Recognition,” Mendeley Data, Version 1, 2022. doi: 10.17632/f45bkkt8pr.1. <https://data.mendeley.com/datasets/f45bkkt8pr/1>
- [7] A. Nahapetyan, S. Prasad, K. Childs, A. Oest, Y. Ladwig, A. Kapravelos, and B. Reaves, “On SMS Phishing Tactics and Infrastructure,” Proceedings of the IEEE Symposium on Security and Privacy, 2024. Dataset: <https://github.com/wspr-ncsu/sms-phishing>
- [8] H. W. Chung et al., “Scaling Instruction-Finetuned Language Models,” arXiv:2210.11416, 2022. <https://arxiv.org/abs/2210.11416>
- [9] Qwen Team, “Qwen2.5-1.5B-Instruct,” Hugging Face Model Card, 2024. <https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct>
- [10] ANZ, “Account Locks and Card Blocks.” <https://www.anz.com.au/support/online-banking/account-management/locked-accounts-card-blocks/>
- [11] ITV, “The Win O’Clock Quiz Prize Draw: Terms and Conditions of Entry.” <https://www.itv.com/terms/articles/the-win-oclock-quiz-prize-draw-terms-and-conditions-of-entry>
- [12] U.S. Bank, “What Are Some Reasons I Might Receive a Security Alert?” <https://www.usbank.com/customer-service/knowledge-base/KB0069455.html>

[13] O2, “Prize Draw and Competition Winners.”

<https://www.o2.co.uk/priority/faqs/prize-draw-and-competition-winners>

[14] Unity Bank, “Protecting Your Accounts with 24/7 Fraud Monitoring.”

<https://www.unity.bank/help/security/fraud-and-security/protecting-your-accounts-with-247-fraud-monitoring/>

[15] A. R. Ahmed and G. F. Islam, “Financial Scams Detection Dataset,” Mendeley Data,

Version 2, 2025. doi: 10.17632/znsk27yk3h.2. <https://data.mendeley.com/datasets/znsk27yk3h/2>

[16] ITV, “Interactive Competitions FAQs.” <https://www.itv.com/terms/articles/interactivefaqs>

[17] Global, “Premium Rate Text & Interactive Services: Frequently Asked Questions.”

<https://global.com/legal/premium-rate-text-interactive-services/>

[18] TD Canada Trust, “Two-Step Verification.”

<https://www.td.com/ca/en/about-td/privacy-and-security/how-we-protect-you/two-step-verification>

[19] Vodafone, “VeryMe Rewards Spring 2025.”

<https://www.vodafone.co.uk/newscentre/press-release/veryme-rewards-spring-2025-spin-to-win-25000quid/>

7. Appendices

Appendix A. Original Data Records

A.1 Original Scam Messages

Table A1 contains the 250 original Scam messages used before data splitting and augmentation. Each row includes the record ID, message text, label, category, source, source URL, source reference, and source type when available.

Link:

https://docs.google.com/spreadsheets/d/158Fa_-smDynaDEQSamW89ltBJlys6YZr/edit?usp=sharing&oid=109675868124521287644&rtpof=true&sd=true

A.2 Original Normal Messages

Table A2 contains the 250 original Normal messages used before data splitting and augmentation. The records appear in the same form used by the data preparation notebook.

Link:

<https://drive.google.com/file/d/1roW5BaNFEWIPQ4E5r49WJTbraRJrEnHj/view?usp=sharing>

The project did not create a separate redacted copy. Readers should not open or contact any link, phone number, or email address shown in these records.

Appendix B. Data Preparation Audit (Augmentation Record)

The table lists each generated message, its parent ID, label, and augmentation method. All augmentation parents came from the original training set.

Link:

<https://drive.google.com/file/d/15mDLO1X3PecSw4eFTNXWJ75INcIY7vxU/view?usp=sharing>

Appendix C. Qwen and Validator Tests

This appendix presents examples from the Qwen explanation tests, including successful explanations for Normal and Scam messages, failed validation cases, and cases where the validator did not detect conflicting information. Each example includes the original message, Qwen's explanation, validation results, and final output when available.

C.1 Successful Explanation — Normal Message

Figure C1 shows a Normal message from the test set (normal_110). The classifier predicted Normal with a scam score of 0.3549, below the threshold of 0.38. Qwen explained the message as a tax refund notification.

The technical record shows that all nine validation checks passed on the first attempt.

Take a closer look.

Message to analyze

English messages. Remove real names, account numbers and codes first.

HMRC: Your requested Self Assessment tax refund is being processed. We will let you know when to expect it.

134 / 2,400 characters

Try an illustrative example

Bank alert

Gift card request

Campus reminder

Code protection

☒ Explain with AI

Qwen runs on this app server.

CLASSIFICATION RESULT

Normal

0.3549 Scam score / 1

Decision threshold: 0.38

Below threshold

This score is close to the decision threshold, Please review the message carefully.

Qwen · AI explanation

The message from HMRC confirms that a Self Assessment tax refund is being processed and will notify the recipient when it can be expected. There are no apparent risks based on the information provided.

Technical record & copyable summary

```
12      ,
13      ▼ "checks": {
14        "nonempty_prose": true,
15        "generation_finished": true,
16        "quotes_match_if_present": true,
17        "numeric_values_in_sms": true,
18        "no_new_billing_unit": true,
19        "no_unqualified_verdict": true,
20        "no_claim_of_verification": true,
21        "no_metric_narration": true,
22        "no_direct_instruction_to_follow_sms": true
23      },
24      "passed": true,
25      "error_type": null,
26      "error_message": null,
27      ▼ "attempts": [
28        ► { Object(4) }
29      ],
30      "model_id": "Qwen/Qwen2.5-1.5B-Instruct",
31      "model_revision": "989aa7980e4cf806f80c7fef2b1adb7bc71aa306",
32      "validation_scope": "Basic checks only; human review is still needed."
33    }
34  }
35}
```

Copyable summary

Sift assessment: Normal

Scam model score: 0.3549

Saved threshold: 0.3800

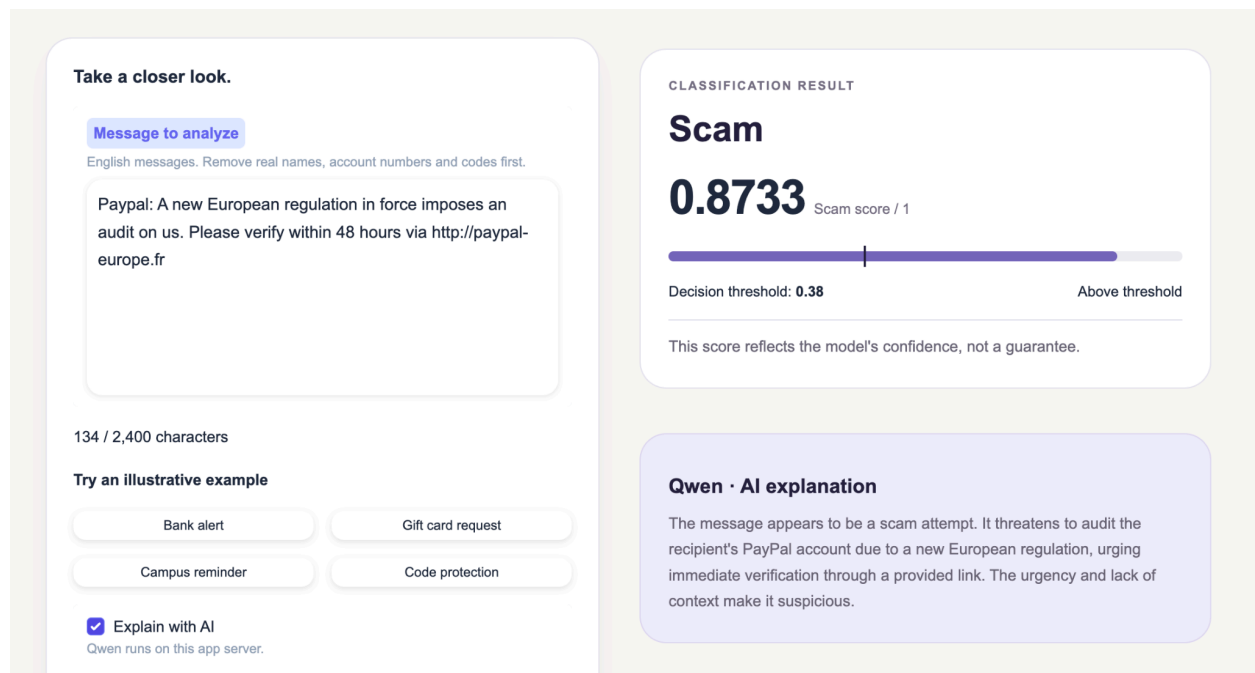
Explanation: Qwen · AI explanation

This result does not verify the sender or guarantee safety.

C.2 Successful Explanation — Scam Message

Figure C2 shows a Scam message from the test set (scam_062). The message claims that PayPal requires an audit and asks the recipient to visit a website within 48 hours. The classifier predicted Scam with a score of 0.8733, above the threshold of 0.38.

The technical record shows that Qwen's explanation passed all nine validation checks on the first attempt.



Technical record & copyable summary

```
12      /
13      ▼ "checks": {
14        "nonempty_prose": true,
15        "generation_finished": true,
16        "quotes_match_if_present": true,
17        "numeric_values_in_sms": true,
18        "no_new_billing_unit": true,
19        "no_unqualified_verdict": true,
20        "no_claim_of_verification": true,
21        "no_metric_narration": true,
22        "no_direct_instruction_to_follow_sms": true
23      },
24      "passed": true,
25      "error_type": null,
26      "error_message": null,
27      ▼ "attempts": [
28        ► { Object(4) }
29      ],
30      "model_id": "Qwen/Qwen2.5-1.5B-Instruct",
31      "model_revision": "989aa7980e4cf806f80c7fef2b1adb7bc71aa306",
32      "validation_scope": "Basic checks only; human review is still needed."
33    }
34  }
```

Copyable summary

Sift assessment: Scam
Scam model score: 0.8733
Saved threshold: 0.3800
Explanation: Qwen · AI explanation
This result does not verify the sender or guarantee safety.

C.3 Failed Validation and Fallback

Figure C3 shows a synthetic message used to test the validation and fallback process. The classifier predicted Scam with a score of 0.7316. However, Qwen repeated the message's instruction to pay \$25, causing the `no_direct_instruction_to_follow_sms` check to fail.

After two failed attempts, the system displayed "AI explanation unavailable" while keeping the original classification result unchanged.

Take a closer look.

Message to analyze

English messages. Remove real names, account numbers and codes first.

The account has been locked. Please pay \$25 today to restore your access.

0 / 2,400 characters

Try an illustrative example

Bank alert

Gift card request

Campus reminder

Code protection

☒ Explain with AI

Qwen runs on this app server.

CLASSIFICATION RESULT

Scam

0.7316

Scam score / 1

Decision threshold: 0.38

Above threshold

This score reflects the model's confidence, not a guarantee.

AI explanation unavailable

No checked AI explanation was produced. Open Technical record for the generated text and error details.

Technical record & copyable summary

Backend record and original-output checks

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

▼ {

"prediction": "Scam",

"predicted_label": 1,

"scam_score": 0.7315994413453917,

"threshold": 0.38,

"near_threshold": false,

"non_latin_text": false,

▼ "explanation": {

"source": "AI explanation unavailable",

"summary":

"No checked AI explanation was produced. Open Technical record for the generated text and error details.",

"raw_output": "The account has been locked. Please pay \$25 today to restore access.",

▼ "checks": {

"nonempty_prose": true,

"generation_finished": true,

"quotes_match_if_present": true,

"numeric_values_in_sms": true,

"no_new_billing_unit": true,

"no_unqualified_verdict": true,

"no_claim_of_verification": true,

"no_metric_narration": true,

"no_direct_instruction_to_follow_sms": false

Copyable summary

Sift assessment: Scam

Scam model score: 0.7316

Saved threshold: 0.3800

Explanation: AI explanation unavailable

This result does not verify the sender or guarantee safety.

C.4 Unresolved Explanation Conflict

Figure C4 shows an example from the original Scam dataset (scam_163). The classifier predicted Scam with a score of 0.9336. Qwen identified the urgent callback request as a pressure tactic but also stated that there was no apparent risk.

This conflicting explanation passed the automatic validation checks. This shows that the validator can detect certain errors but cannot identify every problem in Qwen's reasoning.

Take a closer look.

Message to analyze

English messages. Remove real names, account numbers and codes first.

<Forwarded from 448712404000>Please CALL 08712404000 immediately as there is an urgent message waiting for you.

0 / 2,400 characters

Try an illustrative example

Bank alert

Gift card request

Campus reminder

Code protection

☒ Explain with AI

Qwen runs on this app server.

CLASSIFICATION RESULT

Scam

0.9336

Scam score / 1

Decision threshold: 0.38

Above threshold

This score reflects the model's confidence, not a guarantee.

Qwen · AI explanation

The message forwards from 448712404000 instructs the recipient to call 08712404000 immediately due to an urgent message. This appears to be a pressure tactic with no apparent risk.

Technical record & copyable summary

```
12     },
13     "checks": {
14       "nonempty_prose": true,
15       "generation_finished": true,
16       "quotes_match_if_present": true,
17       "numeric_values_in_sms": true,
18       "no_new_billing_unit": true,
19       "no_unqualified_verdict": true,
20       "no_claim_of_verification": true,
21       "no_metric_narration": true,
22       "no_direct_instruction_to_follow_sms": true
23     },
24     "passed": true,
25     "error_type": null,
26     "error_message": null,
27     "attempts": [
28       { Object(4) }
29     ],
30     "model_id": "Qwen/Qwen2.5-1.5B-Instruct",
31     "model_revision": "989aa7980e4cf806f80c7fef2b1adb7bc71aa306",
32     "validation_scope": "Basic checks only; human review is still needed."
33   }
34 }
```

Copyable summary

Sift assessment: Scam
Scam model score: 0.9336
Saved threshold: 0.3800
Explanation: Qwen · AI explanation
This result does not verify the sender or guarantee safety.