

Explainable Robustness of Multimodal Spam Detection under Adversarial Text Attacks

Samantha Acosta-Ruiz¹, Mireya Tovar-Vidal¹, José A. Reyes-Ortiz²

¹ Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla,
Mexico

² Universidad Autónoma Metropolitana,
División de Ciencias Básicas e Ingeniería, Azcapotzalco, Mexico City,
Mexico

jaro@azc.uam.mx, ar224570157@alm.buap.mx, mireya.tovar@correo.buap.mx

Abstract. Multimodal spam detection systems achieve high performance under clean conditions by combining textual and visual information; however, their robustness under adversarial scenarios remains insufficiently explored. While textual adversarial attacks have been widely studied in text-only classification, it is still unclear whether visual information helps preserve spam detection performance when the textual branch is intentionally perturbed. Results indicate that PWWS-based adversarial attacks significantly reduce spam recall in text-only classification from 0.9820 to 0.8619, thereby increasing critical false negatives. In contrast, multimodal models preserve substantially higher recall under the same attack conditions, reaching 0.9708, which indicates that visual information acts as a robustness support mechanism. Grad-CAM and OCR overlap analysis further support these findings by showing stronger visual alignment in rescue cases and weaker correspondence when both modalities fail. The results suggest that multimodal fusion provides not only better classification performance under clean conditions, but also greater resilience against adversarial manipulation, underscoring the need for adversarial robustness evaluation in practical multimodal spam detection systems.

Keywords: Multimodal spam detection, adversarial text attacks, robustness evaluation, multimodal fusion, explainable AI.

Robustez explicable de la detección multimodal de spam frente a ataques adversarios de texto

Resumen. Los sistemas multimodales de detección de spam logran un alto rendimiento en condiciones normales al combinar información textual y visual; sin embargo, su robustez en escenarios adversarios no ha sido suficientemente explorada. Si bien los ataques adversarios

textuales han sido ampliamente estudiados en la clasificación basada únicamente en texto, aún no está claro si la información visual ayuda a mantener el rendimiento de la detección de spam cuando la rama textual es perturbada intencionalmente. Los resultados indican que los ataques adversarios basados en PWWS reducen significativamente la tasa de recuperación (*recall*) de spam en la clasificación de solo texto, pasando de 0,9820 a 0,8619, lo que aumenta el número de falsos negativos críticos. En cambio, los modelos multimodales mantienen una tasa de recuperación sustancialmente mayor bajo las mismas condiciones de ataque, alcanzando 0,9708; esto demuestra que la información visual actúa como un mecanismo de apoyo a la robustez. El análisis de superposición mediante Grad-CAM y OCR respalda estos hallazgos al mostrar una alineación visual más fuerte en los casos recuperados con éxito y una correspondencia más débil cuando ambas modalidades fallan. Los resultados sugieren que la fusión multimodal no solo proporciona un mejor rendimiento de clasificación en condiciones normales, sino también una mayor resiliencia frente a la manipulación adversaria, lo que subraya la necesidad de evaluar la robustez adversaria en sistemas prácticos de detección multimodal de spam.

Palabras clave: Detección multimodal de spam, ataques adversarios de texto, evaluación de robustez, fusión multimodal, IA explicable.

1 Introduction

The growing reliance on email communication in personal, academic and organizational settings has also increased the exposure to cybersecurity threats. Spam emails remain one of the most common attack vectors for phishing campaigns, malware distribution, fraudulent promotions and social engineering attempts, making reliable spam detection an essential task in modern cybersecurity [3]. To mitigate these risks, machine learning and deep learning models have been widely adopted for automatic spam filtering, often achieving high performance under clean and controlled conditions. In particular, recent comparative studies have shown that transformer-based architectures, such as DistilBERT, provide highly competitive results for malware and spam email identification, highlighting the effectiveness of NLP-based approaches for email security tasks [4].

However, strong performance under normal conditions does not guarantee robustness against adversarial manipulation. Attackers can introduce lexical, syntactic or semantic perturbations that preserve the malicious intent of an email while altering its representation enough to fool the classifier. Even small modifications are usually enough to reduce the detection performance, especially in security-sensitive applications, such as spam filtering, where misclassifying malicious messages as legitimate can allow phishing attempts, malware delivery and other harmful threats to bypass protection mechanisms [10].

In real-world scenarios, spam emails are rarely limited to plain text. Many malicious campaigns rely on embedded images, banners, logos, visual promotions

and structured layouts that complement or even replace textual content to evade traditional text-based filters. In these cases, the malicious signal is frequently distributed across both textual and visual modalities, which makes text-only analysis insufficient for reliable detection. This limitation has driven the development of multimodal spam detection systems that combine textual and visual information, allowing models to capture complementary evidence and improve classification performance [11].

Nevertheless, most adversarial robustness studies still focus on text-only classification or report multimodal results using aggregate metrics such as accuracy and F_1 -Score. These metrics provide limited insight into whether visual information can compensate for adversarial perturbations introduced into textual content and under which conditions multimodal rescue or failure occurs. Understanding these rescue and failure cases is particularly important for determining whether multimodal architectures offer genuine robustness rather than performance gains only under clean conditions.

This work explores the explainable robustness of multimodal spam detection against adversarial text attacks. We investigate whether combining textual and visual information improves robustness when adversarial perturbations are introduced into the textual branch, and whether visual information helps preserve correct spam classification under attack conditions. Through Grad-CAM, OCR overlap analysis, and rescue-case evaluation, we provide a more interpretable framework for understanding adversarial robustness beyond conventional aggregate performance metrics, enabling a more rigorous interpretation of when visual evidence compensates for textual degradation and when both modalities fail.

2 Related Work

The robustness of spam detection systems has become an increasingly important research topic due to the growing sophistication of adversarial manipulation in email communication. Although modern classifiers achieve strong performance under standard evaluation settings, their reliability can be significantly reduced when attackers introduce carefully crafted perturbations designed to evade detection. In natural language processing, this problem is particularly critical because small lexical or contextual modifications are often sufficient to alter model predictions without changing the semantic intent of the original message.

Early adversarial text attacks demonstrated that carefully crafted perturbations can significantly degrade classifier performance while preserving the original semantic intent. TextBugger [6] introduced character- and word-level perturbations capable of deceiving deep text classifiers under both white-box and black-box settings, highlighting the vulnerability of NLP models to subtle textual modifications. Building upon this idea, PWWS [7] showed that identifying influential words through probability-weighted saliency and replacing them with semantically similar alternatives can effectively fool classifiers while maintaining readability.

The effectiveness of saliency-guided adversarial generation was further demonstrated by TextFooler [5], which uses word-importance ranking and synonym substitution to reduce classification performance while preserving semantic consistency substantially. More recently, TextJosher [9] extended these ideas through a transfer-based black-box attack capable of generating adversarial examples that remain effective across different models without requiring direct access to the target classifier.

While adversarial attacks have been studied extensively in text-only classification, real spam campaigns often combine textual and visual content to evade traditional filters. Images, banners, logos and structured document layouts often contain relevant malicious information that cannot be captured through textual analysis alone. This limitation has motivated the development of multimodal spam detection systems that integrate both modalities to improve classification performance.

Architectures such as MMPC-RF [2] and MMTD [11] show that feature-level fusion between textual and visual representations consistently outperforms unimodal approaches, particularly when malicious intent is distributed across both modalities.

Similarly, recent studies using DistilBERT combined with structural and multimodal features report stronger classification performance and greater robustness in spam detection tasks [1]. However, these studies primarily evaluate classification effectiveness under clean conditions and provide limited evidence regarding adversarial robustness.

In parallel, explainability methods have become increasingly relevant for understanding model behavior under adversarial conditions. Techniques such as Grad-CAM [8] make it possible to visualize the image regions that contribute most to the final prediction, helping determine whether the model relies on semantically meaningful evidence or on irrelevant visual artifacts. This capability is particularly important in multimodal security applications because apparent robustness may result from spurious correlations rather than genuine complementary information between modalities. However, explainability is rarely incorporated into adversarial evaluations of multimodal spam detection systems.

Despite advances in adversarial NLP, multimodal spam detection, and explainable AI, their intersection remains insufficiently explored. Existing studies either focus on adversarial attacks in text-only settings or evaluate multimodal spam detection under clean conditions, primarily emphasizing performance improvements achieved through fusion strategies. Consequently, there is limited understanding of how textual adversarial perturbations affect multimodal robustness and whether visual information can compensate for the resulting degradation in spam detection performance.

Furthermore, the contribution of the visual modality as a robustness mechanism under adversarial conditions remains unclear. As a result, it is still uncertain whether multimodal architectures provide genuine robustness against adversarial perturbations or merely achieve improved predictive performance under standard evaluation settings.

3 Methodology

The experiments were carried out using the MMTD dataset [11], a multilingual and multimodal spam detection corpus containing textual content, document images and binary labels for spam and ham classification. The original dataset consists of 31,212 samples organized into three main attributes: textual message (*texts*), associated image path (*pics*) and class label (*labels*), where 1 represents spam and 0 represents ham. Since this work focuses on adversarial robustness in the textual branch, only English samples were selected to ensure linguistic consistency during the adversarial generation. Language identification was performed through the textual modality and the entries without textual content (*no_text*) were eliminated, resulting in 7,712 samples in English.

To improve data quality, text-length distribution was analyzed using boxplot-based quartile filtering to identify and remove atypically short and excessively long messages. This process reduced noise caused by abnormal samples and produced a final textual subset of 7,082 instances while preserving the original class imbalance between spam and ham messages. All retained samples were also verified to maintain valid correspondence with their associated document images, ensuring consistency for multimodal training and evaluation. A standard text normalization process was applied using the *textacy* preprocessing pipeline, including lowercase conversion, URL removal, punctuation normalization, elimination of non-semantic artifacts, and whitespace normalization. This step preserved the semantic content of each message while reducing irrelevant noise for both classification and adversarial evaluation. Since adversarial perturbations are introduced later at the lexical level, preprocessing focused on preserving meaningful textual structure rather than orthographic artifacts.

The dataset was divided using a stratified 70/15/15 strategy for training, validation, and final testing, preserving the original class distribution. The training and validation partitions were used during a 3-fold stratified cross-validation procedure to obtain robust performance estimates and standard deviation analysis, while the final test partition was reserved exclusively for the robustness evaluation. DistilBERT was selected as the text-only baseline due to its strong performance and lower computational cost compared to larger transformer architectures. The maximum sequence length was set to 128 tokens, as the text-length distribution indicated that most messages could be fully represented within this range while maintaining computational efficiency.

For the multimodal model, DistilBERT and ResNet50 were employed as textual and visual encoders, respectively. The textual representation was obtained from the first-token embedding of the final hidden layer of DistilBERT (*last_hidden_state[:,0,:]*), which provides a contextual representation of the complete message. For the visual branch, a ResNet50 model pretrained on ImageNet was used, replacing its original fully connected classification layer with an identity mapping. Consequently, the visual representation corresponded to the 2048-dimensional feature vector produced after the global average pooling stage of the network. To facilitate multimodal fusion, both textual and visual representations were projected into 256-dimensional vectors and concatenated

into a unified 512-dimensional representation, which was subsequently processed through fully connected layers for the final spam classification.

The adopted early-fusion strategy enables the model to jointly learn complementary information from both modalities and evaluate whether visual evidence can compensate for the adversarial degradation introduced into the textual branch.

To evaluate the adversarial robustness, PWWS [7] was applied to the test set as a textual attack method. PWWS generates adversarial examples by identifying the most influential words for classification and replacing them with semantically similar alternatives while preserving readability and malicious intent. Since the attack targets only the textual modality, two adversarial evaluation configurations were considered: text-only attack and multimodal adversarial setting. In the first case, DistilBERT was evaluated directly on the perturbed text. In the second case, the adversarial text replaced the original textual branch, while the visual modality remained unchanged, allowing an analysis of whether visual information helps preserve spam detection performance under attack conditions.

This design intentionally isolates the effect of textual adversarial perturbations, allowing the contribution of visual information to robustness preservation to be evaluated independently from image-based attacks. Consequently, the experimental setup focuses on determining whether visual evidence can compensate for textual degradation rather than evaluating joint multimodal attack strategies.

Because adversarial generation may fail or produce duplicate outputs, a re-alignment stage was required for multimodal evaluation after PWWS generation. While the original multimodal baseline was constructed using the complete aligned text-image test set, adversarial evaluation preserved only samples containing valid adversarial text, document images, and consistent label assignments. After removing duplicates and invalid matches, the final aligned multimodal subset contained 1058 instances composed of 718 spam samples and 340 ham samples. Among them, 171 corresponded to successful adversarial attacks, while 887 retained the status of failed attacks.

Consequently, the clean multimodal subset reported during adversarial evaluation does not correspond to the complete multimodal baseline, but rather to the clean evaluation restricted to the same aligned samples used in the adversarial setting. Therefore, all clean-versus-adversarial comparisons were conducted on the same aligned subset, ensuring a fair and consistent evaluation protocol.

Performance was evaluated using Accuracy, Precision, Recall for the spam class, F_1 -Score, and confusion matrices under both clean and adversarial conditions. Special attention was paid to recall preservation and false-negative behavior since misclassifying spam as legitimate messages represents the highest security risk in practical deployment. In addition, statistical significance testing was performed to determine whether the observed differences between clean and adversarial conditions were attributable to adversarial perturbations rather than random variation.

Table 1. Baseline performance under clean conditions for multi-modal, text-only and image-only spam detection models.

Model	Accuracy	Precision	Recall	Spam F ₁ -score
Text-only (DistilBERT)	0.9793	0.9875	0.9820	0.9848
Image-only (ResNet50)	0.9566	0.9827	0.9528	0.9675
Multimodal (Text + Image)	0.9929	0.9927	0.9969	0.9948

Finally, an explainability analysis was incorporated to investigate rescue and failure behavior under adversarial conditions. Grad-CAM was used to visualize the image regions that most contributed to spam predictions, while OCR overlap was quantified using the Intersection over Union (IoU) between OCR-extracted text regions and Grad-CAM activation areas. This analysis provided both qualitative and quantitative evidence regarding whether visual attention remained aligned with semantically relevant spam content and whether the visual modality contributed to preserving correct classifications under textual adversarial perturbations.

4 Results

This section presents the experimental evaluation of multimodal spam detection under textual adversarial attacks. The analysis focuses on whether visual information helps preserve spam detection performance when the textual branch is intentionally manipulated, with special attention to recall preservation for the spam class, since false negatives represent the highest security risk in practical spam filtering systems. In addition to aggregate metrics, Grad-CAM and OCR overlap are used to analyze rescue and failure cases under adversarial conditions.

Following the experimental protocol described in the methodology section, model performance was evaluated using Accuracy, Precision, Recall for the spam class, and F₁-score. A 3-fold stratified cross-validation procedure was applied over the development split to ensure reliable performance estimation, while final results were obtained using the independent test set.

Table 1 summarizes the performance of the model in clean conditions for text-only, image-only and multimodal configurations. DistilBERT achieves strong results in text-only classification, while ResNet50 also shows competitive performance using only document images. However, the multimodal model achieves the best overall performance, reaching the highest Accuracy, Precision, Recall, and F₁-score. In particular, the recall of spam increases from 0.9820 in the text-only configuration to 0.9969 in the multimodal model, suggesting that the textual and visual modalities provide complementary information for spam detection. These results establish a solid baseline and suggest that multimodal fusion improves detection reliability before introducing adversarial perturbations. This baseline serves as a reference point to quantify the impact of adversarial perturbations and assess whether the observed differences are still statistically significant under attack conditions.

Table 2. Performance degradation in the text-only model under PWWS adversarial attack.

Setting	Accuracy	Precision	Recall Spam	F ₁ -score
Text-only clean	0.9793	0.9875	0.9820	0.9848
Text-only attacked (PWWS)	0.8383	0.8966	0.8619	0.8789

Table 3. Multimodal performance under clean and adversarial textual conditions evaluated on the same test partition.

Setting	Accuracy	Recall Spam	F ₁ -score
Multimodal clean	0.9339	0.9654	0.9519
Multimodal adversarial	0.8222	0.9434	0.8780

To evaluate whether this robustness remains under adversarial manipulation, PWWS was first applied to the text-only setting using DistilBERT. Since the attack targets only the textual modality, this experiment allows for the quantification of the impact of lexical perturbations on spam detection performance before analyzing the multimodal setting.

Table 2 shows a substantial degradation in performance under attack, with Accuracy decreasing from 0.9793 to 0.8383 and Recall for the spam class dropping from 0.9820 to 0.8619. This corresponds to an absolute decrease of 14.1 percentage points in Accuracy and 12.0 percentage points in spam Recall. This reduction indicates that lexical perturbations are sufficient to alter model predictions while preserving semantic meaning, increasing false negatives and exposing the vulnerability of text-only spam detection systems under adversarial conditions. These findings suggest that relying solely on textual information makes the classifier highly sensitive to adversarial manipulation.

After establishing the vulnerability of the text-only model, the next step was to analyze whether multimodal fusion helps preserve performance when the textual branch is attacked. To ensure a fair comparison between clean and adversarial conditions, PWWS perturbations were generated directly from the same multimodal test partition. Consequently, each adversarial sample preserved its original image, label, and sample identifier, allowing clean and adversarial evaluations to be performed on exactly the same set of 1,406 instances. Therefore, the multimodal results reported in Table 3 correspond to paired evaluations in which only the textual modality was modified while the visual information remained unchanged. This setup enables a direct assessment of the extent to which visual information contributes to maintaining spam detection performance under adversarial textual perturbations.

Table 3 presents multimodal performance under textual adversarial attack. Although the overall Accuracy decreases from 0.9339 to 0.8222 and Recall for the spam class decreases from 0.9654 to 0.9434, this corresponds to an absolute reduction of 11.2 percentage points in Accuracy but only 2.2 percentage points in spam Recall. These results suggest that the visual modality helps preserve spam detection sensitivity even when the textual branch is compromised. Unlike the text-only setting, where adversarial perturbations produce substantially larger

Table 4. Comparison between text-only and multimodal models under adversarial conditions.

Model	Accuracy	Recall Spam	F ₁ -score
Text-only attacked	0.8383	0.8619	0.8789
Multimodal adversarial	0.8222	0.9434	0.8780

Table 5. Statistical significance analysis of spam recall degradation under adversarial conditions.

Model	Recall Clean	Recall Attack	Z-score	P-value
Text-only	0.9820	0.8619	7.34	2.13×10^{-13}
Multimodal	0.9654	0.9434	2.30	0.0212

degradation, multimodal fusion provides complementary evidence that partially compensates for lexical manipulation. To determine whether these differences were statistically meaningful rather than the result of random variation, a statistical significance analysis was subsequently performed.

To directly compare both attacked settings, Table 4 contrasts the performance of the text-only attacked model and the multimodal adversarial model. This comparison is particularly important because it allows us to determine whether visual information functions only as complementary support under clean conditions or as a genuine robustness mechanism under adversarial attacks.

As shown in Table 4, while the text-only attacked model achieves a spam recall of 0.8619, the multimodal adversarial model preserves a substantially higher recall of 0.9434 under the same attack conditions. This corresponds to an absolute improvement of 8.2 percentage points in spam recall. Although the multimodal model exhibits slightly lower Accuracy, it maintains considerably stronger spam detection capability, as reflected by its higher recall and lower number of false negatives. Specifically, the number of spam emails incorrectly classified as legitimate messages decreases from 92 in the text-only setting to 54 in the multimodal configuration. These results suggest that visual information provides complementary evidence that helps compensate for the degradation introduced in the textual modality.

Although the previous results indicate that multimodal fusion preserves substantially higher spam recall under adversarial conditions, it is important to determine whether the observed differences are statistically significant rather than the result of random variation. Therefore, a two-proportion Z-test was conducted using spam recall under clean and adversarial conditions for both the text-only and multimodal models.

Table 5 shows that the degradation produced by PWWS is statistically significant for both architectures ($p < 0.05$). However, the impact is considerably smaller in the multimodal setting. While the text-only classifier experiences a reduction of 12.0 percentage points in spam recall, the multimodal model shows a decrease of only 2.2 percentage points. Despite the adversarial perturbations, the multimodal architecture preserves a substantially higher spam detection capability. These findings provide statistical evidence that multimodal fusion

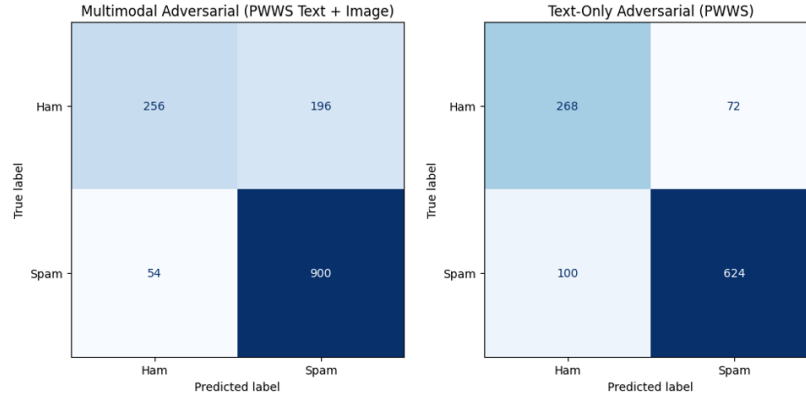


Fig. 1. Confusion matrices for the text-only and multimodal models under PWWS adversarial conditions.

reduces the impact of textual adversarial attacks and support the hypothesis that visual information contributes to maintaining spam detection performance when the textual modality is compromised.

While the statistical analysis quantifies the significance of the observed degradation, it does not reveal how the errors are distributed across classes. To further examine the practical impact of adversarial perturbations, Figure 1 presents the confusion matrices obtained for the text-only and multimodal models under adversarial conditions. It should be noted that the confusion matrices correspond to the evaluation settings used for each experiment and therefore contain different class distributions. Consequently, absolute counts should not be directly compared across models. Instead, the analysis focuses on the relative distribution of classification errors, particularly false negatives, since these determine the spam recall and represent the most critical failure in spam filtering systems.

As shown in Figure 1, the text-only model incorrectly classifies 100 spam emails as legitimate messages, whereas the multimodal model reduces this number to 54. Although the absolute counts originate from evaluation sets with different class distributions, the reduction in false negatives is consistent with the higher spam recall achieved by the multimodal architecture. From a cybersecurity perspective, this behavior is particularly important because false negatives allow malicious emails to bypass the filtering system. Overall, the confusion matrix analysis confirms that multimodal fusion better preserves spam detection performance under adversarial conditions.

Although the above analyses demonstrate that multimodal fusion improves robustness against textual adversarial attacks, they do not explain how the visual modality contributes to the final decision process.

To gain further insight into this behavior, an explainability analysis based on the superposition of Grad-CAM and OCR was performed. Specifically, the Intersection over Union (IoU) metric was used to quantify the overlap between

Table 6. OCR-GradCAM overlap measured using IoU under clean and adversarial conditions.

Setting	Mean IoU	Std IoU	Median IoU	Valid Samples
Multimodal Clean	0.0744	0.0556	0.0705	1406
Multimodal Adversarial	0.0686	0.0545	0.0631	1406

Table 7. Rescue and failure case analysis under adversarial conditions.

Case Type	Count	Percentage (%)	Mean IoU	Std IoU
Both Correct	1027	73.04	0.0708	0.0546
Both Incorrect	113	8.04	0.0635	0.0578
Multimodal Failure	137	9.74	0.0577	0.0539
Rescue Case	129	9.17	0.0669	0.0502

Grad-CAM activation regions and text regions detected by OCR under clean and adversarial conditions.

Table 6 presents the overlap between Grad-CAM activation regions and OCR-detected text areas for clean and adversarial inputs. Although PWWS alters the textual content, the decrease in overlap remains relatively small, with the mean IoU dropping from 0.0744 to 0.0686 and the median IoU from 0.0705 to 0.0631. These findings suggest that adversarial perturbations primarily affect the textual branch, while the visual attention patterns remain comparatively stable. Consequently, the visual modality appears to retain access to relevant document information that may help mitigate the impact of textual manipulation.

While IoU provides a global measure of attention stability, it does not reveal how the interaction between textual and visual information influences individual prediction outcomes. Therefore, a rescue and failure analysis was performed to identify situations in which multimodal fusion successfully compensated for errors introduced by adversarial perturbations.

Table 7 presents the distribution of prediction outcomes under adversarial conditions. Most samples (73.04%) were correctly classified by both models, indicating that many adversarial perturbations were insufficient to alter the final prediction. More importantly, 129 samples (9.17%) correspond to rescue cases, where the text-only model produced an incorrect prediction while the multimodal model remained correct. These cases provide direct evidence that visual information can compensate for degradation introduced in the textual modality. In contrast, multimodal failures account for 9.74% of the samples, indicating that visual information does not always improve prediction outcomes. Nevertheless, the presence of a substantial number of rescue cases supports the higher spam recall observed in the multimodal setting and reinforces the hypothesis that visual information contributes positively to robustness under adversarial conditions.

An analysis of the IoU values across case categories reveals that multimodal failures exhibit the lowest average overlap (0.0577), whereas correctly classified samples show higher overlap values (0.0708). Rescue cases present an intermediate overlap value (0.0669). Although these observations do not establish a causal

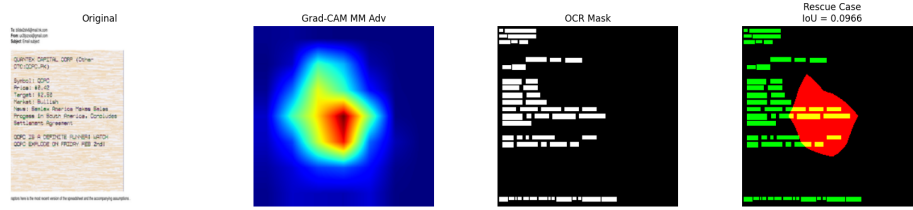


Fig. 2. Grad-CAM and OCR overlap analysis for a rescue case under textual adversarial attack.



Fig. 3. Grad-CAM and OCR overlap analysis for a both-fail case under textual adversarial attack.

relationship, they suggest that stronger alignment between Grad-CAM activations and OCR-detected regions may be associated with more reliable multimodal decisions. To complement this quantitative analysis, representative Grad-CAM visualizations were inspected for rescue and failure cases, providing a qualitative perspective on the model's behavior under adversarial perturbations.

Figure 2 presents a rescue case where the textual branch is successfully attacked, but the multimodal model preserves correct spam classification by maintaining attention over promotional content, suspicious layouts, and highlighted textual areas. The stronger OCR overlap ($\text{IoU} = 0.0966$) indicates better alignment between visual attention and relevant document regions, allowing the visual modality to compensate for textual degradation.

In contrast, Figure 3 illustrates a both-fail case, where neither modality provides sufficient evidence for correct classification. In this example, Grad-CAM exhibits weaker correspondence with relevant document regions, and the lower OCR overlap ($\text{IoU} = 0.0239$) reflects reduced alignment between visual attention and the information contained in the email. Although absolute IoU values are relatively low, this behavior is expected because Grad-CAM provides coarse localization rather than precise segmentation masks. Therefore, IoU should be interpreted as a comparative measure of attention alignment rather than an exact localization metric. The contrast between rescue and failure cases nevertheless helps explain why multimodal systems preserve higher spam recall under adversarial conditions while still exhibiting occasional failures.

Overall, the results consistently indicate that multimodal spam detection provides greater robustness against textual adversarial attacks than text-only classification. While PWWS significantly degrades the performance of the textual

branch, the integration of visual information helps preserve spam recall and reduces critical false negatives under attack. The statistical analysis, confusion matrix evaluation, rescue-case analysis, and Grad-CAM visualizations collectively support the conclusion that visual information contributes to maintaining classification performance when the textual modality is compromised. These findings suggest that multimodal architectures not only improve performance under clean conditions but also provide a more resilient framework for practical spam detection when adversaries attempt to manipulate the textual modality.

5 Conclusions

This work investigated the robustness of multimodal spam detection systems against textual adversarial attacks by comparing text-only and multimodal architectures within a unified experimental framework. Unlike conventional evaluations performed exclusively under clean conditions, this study examined whether visual information can help preserve spam detection performance when the textual modality is intentionally manipulated through PWWS adversarial perturbations.

The experimental results demonstrate that textual adversarial attacks substantially degrade the performance of text-only classification models. In contrast, the multimodal architecture exhibited a considerably smaller reduction in spam recall and a lower number of false negatives, preserving its ability to correctly identify malicious emails under attack. Statistical significance analysis confirmed that both architectures were affected by adversarial perturbations; however, the impact was notably smaller in the multimodal setting. This behavior was further supported by the confusion matrix analysis, which showed a consistent reduction in critical classification errors when visual information was incorporated.

Additional insight was obtained through the explainability analysis. The OCR-GradCAM overlap remained relatively stable under adversarial conditions, while the rescue-case analysis revealed multiple instances in which the multimodal model maintained correct predictions despite failures in the textual branch. Furthermore, representative Grad-CAM visualizations showed that successful predictions tended to preserve attention over visually relevant spam-related regions, whereas failure cases exhibited weaker correspondence between attention maps and document content.

Taken together, these findings indicate that visual information contributes not only to improving classification performance but also to preserving detection capability when the textual modality is compromised. Consequently, evaluating spam detection systems solely under clean conditions may overestimate their practical robustness. The results support multimodal architectures as a more resilient alternative for real-world spam filtering scenarios in which adversaries actively manipulate textual content.

Future work will investigate more challenging multimodal adversarial scenarios, including joint text-image attacks designed to simultaneously

manipulate both modalities. Such experiments would help determine whether the robustness observed in this study remains when visual information is also targeted. Another promising direction involves causal and counterfactual analyses aimed at understanding which visual characteristics are responsible for the rescue behavior observed in multimodal predictions. Additionally, future research will explore multimodal defense strategies, including adversarially robust training and cross-modal consistency mechanisms, to further improve the security and reliability of practical spam detection systems.

References

1. Asliyukse, H., Tonkal, O., Kocaoglu, R.: A comparative evaluation of a multimodal approach for spam email classification using distilbert and structural features. *Electronics* 14(19), 3855 (2025)
2. Hnini, G., Riffi, J., Mahraz, M.A., Yahyaouy, A., Tairi, H.: Mmpc-rf: a deep multimodal feature-level fusion architecture for hybrid spam e-mail detection. *Applied Sciences* 11(24), 11968 (2021)
3. Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., Alegre, E.: A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review* 56(2), 1145–1173 (2023)
4. Jáñez-Martino, F., Fidalgo, E., Alaiz-Rodríguez, R., Carofilis, A., Martínez-Mendoza, A.: Comparative analysis of natural language processing models for malware spam email identification. In: *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*. pp. 59–63 (2024)
5. Jin, D., Jin, Z., Zhou, J.T., Szolovits, P.: Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 34, pp. 8018–8025 (2020)
6. Li, J., Ji, S., Du, T., Li, B., Wang, T.: Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271* (2018)
7. Ren, S., Deng, Y., He, K., Che, W.: Generating natural language adversarial examples through probability weighted word saliency. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. pp. 1085–1097 (2019)
8. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision* 128(2), 336–359 (2020)
9. Wang, P., Zhang, S., Xu, Y., He, X., Xu, W.: Textjoshier: A transfer-based black-box attack method against text classifiers. *Information Sciences* p. 122888 (2025)
10. Wu, B., Zhu, Z., Liu, L., Liu, Q., He, Z., Lyu, S.: Attacks in adversarial machine learning: A systematic survey from the lifecycle perspective. *International Journal of Computer Vision* 134(5), 197 (2026)
11. Zhang, Z., Deng, Z., Zhang, W., Bu, L.: Mmtd: A multilingual and multimodal spam detection model combining text and document images. *Applied Sciences* 13(21), 11783 (2023)