International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Explainable AI (XAI) for Spam Detection: The "Lens" Approach

Raghul Sachin R[1], Arun Karthik V[2], Dr. S. Niveditha[3]

[1,2,3]Department of CSE, SRM Institute of Science & Technology Vadapalani campus, Chennai, India

Abstract

This paper presents a methodology for transparent spam detection based on the use of a Random Forest classifier and Explainable Artificial Intelligence (XAI). We first confirm the classifier's high predictive accuracy and then address the issue of black-box opacity by implementing Local Interpretable Model-agnostic Explanations (LIME) for semantic and contextual transparency. Our method highlights the major issue of false positives, which is quite challenging in the case of machine learning models. Our framework starts by training a Random Forest to identify the raveled textual distribution of a standard SMS and email spam dataset, balanced via SMOTE. The heart of our self-adaptive method is a Human-in-the-Loop (HITL) override feature within this transparent decision space. Giving the models reasoning, which identifies the precise linguistic cues leading to a spam prediction, helps to find the best context-specific whitelist rule. This way not only wrongly labeled emails are recovered but also the filter learns therefore this is an excellent application of explainable models for trustworthy and real cybersecurity.

Keywords: Explainable AI, LIME, Spam Detection, Human-in-the-Loop, Random Forest, Cybersecurity

INTRODUCTION

The explosion of digital communication and our world's connectedness have brought cyber security to a whole new level, for example talks about high-fidelity threat detection models that can fight against spam threats which are getting more and more dangerous [1]. Mostly, in the context of discriminative modeling, the focus was on training models only to identify labeled texts and categorize them with separate labels consequently these models were able to quickly label millions of messages either as safe or dangerous.

However, the emergence of complex, non-linear machine learning frameworks completely changed the situation. These original methods were able to identify data patterns without using the backing of simple heuristic rules, as they looked at detection as deep pattern recognition. At the core of such a system is the continuous optimization over a constantly increasing number of features, which aims not only to approximate the true distribution of texts but at the same time to be able to discriminate between malicious and normal samples in the data. The mathematical foundation of this work is such that it enables a model to very accurately memorize patterns, which in turn allows it to achieve nearly perfect results on standard datasets.

Even though these early machine learning models had the potential to predict quite well, their very first versions had a huge architectural limitation: the "black-box" issue. Because the working of these

systems was completely hidden, there emerged an urgent need for interpretable machine learning that could provide a straightforward and logical explanation of the decisions made automatically [2]. As soon as AI systems were implemented in the real world, certain issues were quickly recognized. To illustrate, the unexplainable False Positives generated by the systems led to the discarding of significant, high-stakes emails without any explanation.

Thorough evaluation of the performance of machine learning algorithms for detecting spam revealed that simply maximizing the predictive accuracy of the model is insufficient if the system does not manage to induce the user's trust [3]. The "Trust Gap" was a significant reason for the decline in the real-world effectiveness of automated filters. It shows us that besides merely predicting, the algorithms also have to give the reason for their predictions. This lack of transparency was mostly resolved with the arrival of Explainable Artificial Intelligence (XAI) and sophisticated tools like Local Interpretable Model-agnostic Explanations (LIME) [4]. Firstly, by launching cross-examining toolkits backed up by a diagnostic framework - local surrogate models, e.g. LIME - a reliable way was established for scrutinizing these black box classifiers. Much more, it indicated that a highly trained classifier is not simply a memorizing machine; on the contrary, it is a complex, hierarchical feature space that it learns. Linguistic vectors in this feature space correspond to semantic attributes that may have different interpretations, e.g. urgency financial coercion, or identity in the case of phishing.

The identification of a meaningful semantic mapping indicated that interpretable models might have applications well beyond the simple binary classification. Among the explainable capabilities, one of the most interesting and practical challenges is the interactive threat mitigation application. This type of work is seeking the reestablishment of belief and comprehension in a system for message filtering, which means that the AI is expected to have not only a semantic transparency so deep that it can explain in detail the reason why an email was considered spam, but also that it stops in a simple silent mode filtering only.

As the area gained more ground, researchers started to pay more attention to the fact that explanations should not only be mathematically correct when considered separately from the user experience, but should also be really helpful. A technology combining Human-in-the-Loop (HITL) machine learning, which stands as our main point of departure for the present work, has been described as a unique and highly effective method. By analogy, it was suggested that a highly trained, perfectly working classifier could be used not as a static, uncompromising filter, but more like a responsive, interactive agent.

By utilizing the priors learned by the Random Forest classifier adjusted with the Synthetic Minority Over-sampling Technique (SMOTE), we demonstrate that critically important semantic, context-aware communication restoration is achievable. This method not only holds the high accuracy of the latest discriminative models but also fundamentally places the issue of user trust at its core, hence it is a great demonstration of the practicality of XAI for personalized cybersecurity.

Related Works

This section surveys the existing literature that contains the baseline of the original research as well as the most recent advancements, which outline the research context of the transparent cybersecurity framework we put forward. First, we investigated the automatic text classification, basing the discussion on the traditional Machine Learning algorithms that have been the standard architectures for classifying high-dimensional text data. Then, we delved into the technical progress of Explainable AI (XAI) in this area. We follow the evolution of predictive models, which were initially black-boxes, becoming

nowadays state-of-the-art ones where we mostly point out Human-in-the-Loop (HITL) optimization strategy, that serves the basis of our proposed method.

A. *Specialized Architectures for Interpretability and Data Collections*

Deep predictive modeling progressed over time, and the research focus gradually shifted beyond simple heuristic rule sets. Several researchers at some point found that their normal clas-sification filters would no longer be sufficient to effectively handle opaque decision boundaries. They mostly noticed the lack of trust in the outputs. One of the main architectural breakthroughs was achieved through the launch of a single, comprehensive way of interpreting model predictions [5]. Basic classifiers are simply filters that carry out the same operation whether or not the user understands the logic. On the other hand, unified explanatory frameworks allow the network to learn a dynamic and learning-based feature selection mode in each channel. In this manner, the network can automatically identify and disregard irrelevant or corrupted information, thereby substantially increasing the level of precision in producing transparent reasons for the outputs.

The addition of detailed concepts, taxonomies, and ways to move towards responsible AI [6] along with providing the exact local details, thoroughly transformed how models use global classification information. Conventional neural networks can understand a very limited range of information, and hence, they can only make a binary decision. By leveraging Formalized Explainable AI (XAI) taxonomies, the model can distinguish not only surface features but also the structural patterns of the decision space. Having this kind of global knowledge is especially crucial for the achievement of symmetry and logical consistency in the most difficult situations. Besides, the initial datasets were conceptualized to counteract the mathematical bias that occurs when standard filters are applied to unstructured text. Using a specific set of SMS spam filtering data [7], the researchers are able to be filled with confident vitality that the newly created analytical content will invariably be grounded in genuine, verifiable contexts.

B. *Interpretability, Feature Extraction, and Systematic Reviews*

Increasing the depth of predictive models undoubtedly adds to the research community's predicament of black-box models: how a very high dimensional text vector can be logically consistently classified. Research on the detection of SMS spam messages through term frequency-inverse document frequency (TF-IDF) and Random Forest algorithms has demonstrated a con-crete paper trail for exposing and understanding the internal workings of the networks. Finding these semantic units in a classifier, for example, different explicit keywords, urgent phrases, or contextual inconsistencies, gives one a reason to check the box that yes, the model is indeed making real structural decisions. Such an interpretability is one of the first features of semantic threat detection because it is through it that the system is understood not only to predict the label but to reconstruct the reasoning from the learned linguistic priors.

One of the main outcomes of applying such insights for explain-ability has been recognized in comprehensive systematic literature reviews on the detection and categorization of spam content [9]. Usually, ambiguous cases in unsupervised or black-box filtering are likely to be situations where a flagged message has multiple valid interpretations, and one of these interpretations might be completely safe. Systematic reviews of these methods indicate that feature weighting and extraction limitations provide additional hints to the classification process that eventually lead the predictive engine to a particular structural result. Therefore, the part of the text that is classified normally will be aligning with the user's inbox reasoning stepping in. Here, we are only changing the notion of structural

interpretability so that our message selections are not only visually plausible on a heatmap but will also be semantically accurate in the textual manifold context.

C. *The Paradigm Shift: Trustworthy AI and Human-in-the-Loop*

The threat classification field has been given a fresh boost effect as we know by a few fundamentally new mathematical paradigms that, among other things, discards the reliance on stationary or static accuracy metrics. A great one among these changes is the transition to defining what we have figured out and what is still to be trusted knowledge in the context of artificial intelligence [10]. Machine learning per se is very much performance-oriented; nevertheless, by turning the evaluation into a trust-based realm, the user-centric receptive field can be present at the very first layer even without a user. Thus a model is enabled to consider the overall consequence of a False Positive simultaneously, which is why it can be very large-scale network deployment scenarios where the network needs to handle a lot of private data and still maintain global user trust.

On the contrary, the winning of interactive modes have been really extended to different places, such as cyber security. Human-in-the-loop (HITL) machine learning [11], on the other hand, stands for a major breakthrough in the field of automated filtering. Using human input directly, HITL understands lengthy context dependencies over the whole user journey in a very comprehensive manner. Actually, HITL is making spam filtering an interactive sequence-to-sequence problem, to the extent that the model is able to generate even the more complicated and non-repetitive whitelist rules that pure rigid networks have only been able to dream about until now. The fact that Interactive systems allow user context to 'weigh in' on each and every other baseline rule is what makes them capable of contextual understanding that traditional classifiers cannot go beyond.

D. *Foundations of Stability: Cybersecurity Surveys and Rigorous Science*

The above major achievement in applications in fact, can be backtracked to alterations that have considerably upgraded the functioning of XAI training dynamics and the robustness of these models at their very core. Initially, automated blocking was very unstable mainly due to the fact that it was the only method to achieve high false-positive rates and this problem was resolved by the comprehensive surveys on explainable artificial intelligence for cybersecurity [12]. Enforcing a clear taxonomy on the explaining of threats, these surveys offer a more straightforward implementation scene. In other words, the developer is able to obtain sound and typical insights even when the threat landscape becomes very complicated, which significantly contributes to the stability and variety of the defensive solutions.

On the other hand, striving for full transparency resulted in the discovery of a strict science of interpretable machine learning [13]. This transition has produced a scientific paradigm that can separate the accuracy of the model at a high level from the random learning of humans. A completely different evaluation space of this kind is indispensable for the transparent filtering because it makes it possible for the optimization procedure to hardly change the main structural characteristic of the classifier while the fine stylistic features of the explanation need not be altered. The things leaving us behind is that the architectures have changed from simple classifiers to completely interpretable frameworks that signal the transition from merely matching words to semantically understanding the user's intent.

E. *Rigorous Evaluation and Deep Learning Architectures*

In order to have a fair comparison of these techniques, the scientific community moved away from simple baseline methods, as these methods produce results that are biased towards high accuracy losing explainability. Therefore, advanced deep learning applications for SMS spam filtering [14] have become

the new standard, where the feature space of a pre-trained network is used to determine the statistical distance between the distribu-tions of safe and malicious text. However, acknowledging that a single algorithm is hardly capable of representing both the quality and the transparency needed, researchers employ these deep architectures to execute multi-dimensional evaluations. Such advanced evaluations render quantitative precision and realism of individual text samples possible, thus indicating how well the model covers the entire diversity of the commu-nications domain.

F. *Future Horizons: Hybrid Approaches and Cross-Platform Applications*

Firstly, to be very honest, the best methods for the classification of textual threats are on the threshold of a profound change again, in fact, there is a new set of models that focus on cross-platform behavioral analysis and have emerged as potent challengers of the standalone email filters. The original methods such as those in the surveys of new approaches and the comparative study for Twitter spam detection [15] have turned the filtering procedure on its head by treating the problem as a multi-modal chain in which the gradually removed noise is being conditioned on the known social behaviors. These models not only provide a great capability to support security, but also it is possible for them to create a wide variety of different security profiles for only one user.

Hybrid models are certainly the primary focus of research nowadays however the XAI-based Random Forest has been consistently demonstrating that it is still a very efficient and interpretable method, requiring only a moderate amount of computational resources, especially when dealing with structured datasets, such as a corpus of SMS. Simply put, we are focusing on turning this very basic method into an executable one and performed a complete check of the implemented version, which we then provide to the users along with a clean, reproducible pipeline covering everything from text vectorization through to semantic override. We initially revealed that very distinct research areas trusted AI frameworks and human-in-the-loop stability, for instance, that transparent optimization is a potent and effective tool that can be utilised to address cybersecurity issues.

MATERIALS AND METHODS

To realize our transparent spam detection framework, we went through a series of quantitatively documented steps. First, we gathered text data, then we explained what structural components our natural language processing pipeline had, next we went through the predictive training processes, and finally, we introduced the Human-in-the-Loop (HITL) method for semantic override. A Random Forest classifier pre-trained to a prior in combination with Explainable AI allows the system to regain trust in automated filters with very high semantic correspondence.

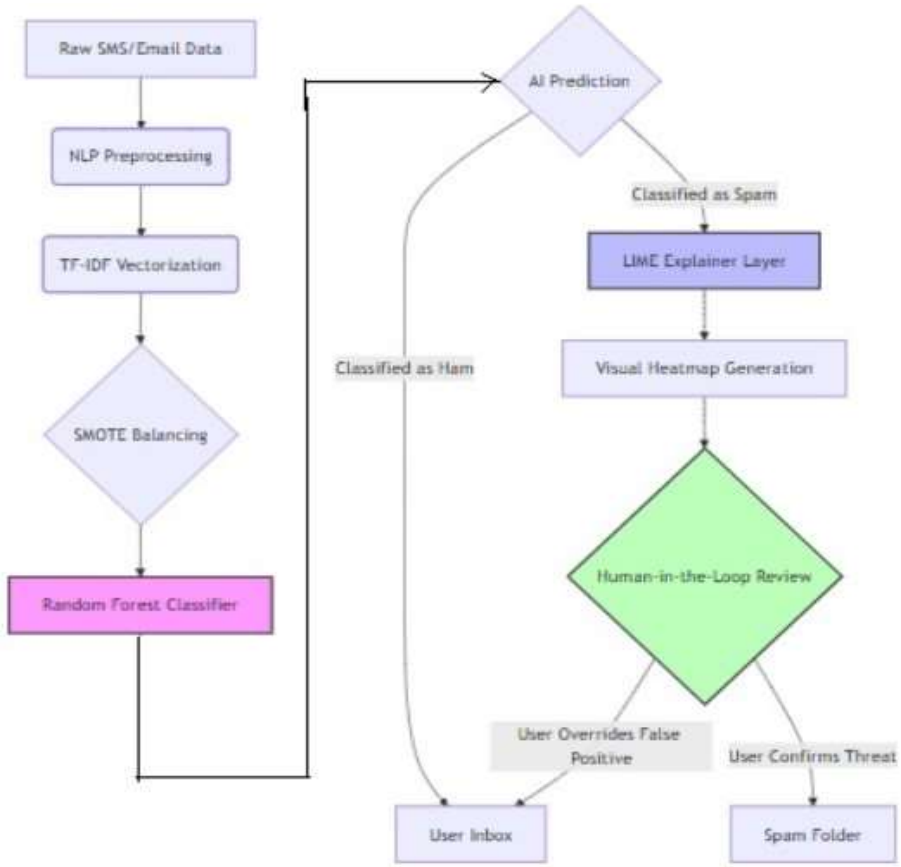

Fig. 1. Architecture Diagram of SMOTE and Random Forest Classifier Pipeline.

A. Dataset and Pre-processing

The work considers standard SMS and email datasets that feature examples of both legitimate and malicious communications. Actually, these are the very collections most famous datasets used in natural language processing and cybersecurity research for benchmarking purposes. It is composed of thousands of text samples that are greatly unbalanced corresponding to the training and test split, where legitimate messages (Ham) far outnumber Spam (about 87% to 13%). Each instance in the dataset is an unprocessed text string of different length. The text is always parsed to remove only the main message body, hence a perfect method for evaluating predictive quality. To ensure computational efficiency and mathematical stability during the training of the mod-els, the raw text data were subjected to two key pre-processing steps. First, the unstructured string arrays are changed through a powerful Natural Language Processing (NLP) pipeline that includes tokenization, stop-word removal, and NLTK lemmatization. The cleaning is necessary because the models are made of mathematical architectures, which need input data in a normalized, noise-free format.

The cleaned text strings, which were originally arrays of individual words, are transformed into numerical feature vectors by a Term Frequency-Inverse Document Frequency (TF-IDF) transformation after that. Such kind of vectorization is very important for statistical training, because it not only modifies data distribution according to term importance but also does it in the way that very frequent, non-informative words are not very highly weighted. When such numerical scales are used to represent linguistic importance, classification models do not suffer from the first-stage training instabilities and contextual blind spots, which are quite frequent in the raw bag of words methods.

B. *Machine Learning Architecture and Balancing*

The basic architecture of the framework revolves around a predictive engine that has two stages or components to it—namely, Synthetic Minority Over-sampling Technique (SMOTE) and Random Forest Classifier. The two components are designed to be used in tandem with each other to attain a predictive equilibrium wherein the classifier can identify malicious examples without any bias towards the majority Ham distribution.

Data Balancing Design (SMOTE): The SMOTE module has been designed to perform a transformation in an unbalanced space. The first input to SMOTE is the vectorized form of the TF-IDF training set provided to it. The vectors corresponding to the minority class are then analyzed to generate synthetic examples through interpolation between nearest neighbours. The rationale behind this operation is that the function does not permit the classifier to ignore the minority class by providing it with a balanced dataset containing 50 percent examples each from the minority and majority class; thus, the learning process becomes equitable. The output of SMOTE is then a perfectly balanced feature array.

Random Forest Design: The design of the Random Forest classifier is an ensemble classifier that is fed the balanced high-dimensional vector and returns a binary score as a prediction. The design is made up of a collection of decision trees that operate individually in reducing the variance of the model using a technique called bootstrap aggregating or bagging. However, in order to prevent the classifier from becoming overfitted to the training data too quickly and thus robbing the system of a generalized learning ability, a degree of feature randomness is included in the model. Here, a majority voting mechanism is used in the final layer to produce a conclusive classification based on a level of certainty that a particular input is a genuine threat from a data distribution point of view.

C. *Training Protocol and Validation*

The training phase means that all ensemble methods are wrapped into one single optimization process. Individual decision trees are expected to separate real and spam samples perfectly. They are rated on how well they reduce impurity at a node using the Gini index. Yet, all these individual decision trees are combined into a very strong ensemble method called a Random Forest. Essentially, ensemble methods are pushing individual decision trees to mutually correct each other's contextual errors. A very thorough 5-Fold Cross-Validation technique is employed to back up all training procedures. Besides that, precision, recall, and F1 scores come to the rescue of the evaluation metric being used instead of accuracy to make sure the model considers all the possible effects of misclassification. Besides, to ensure that the classifier is not too biased towards a single dominating feature that ends up being the reason for all decision-making processes and, thus, creating a situation where blind spots are formed, a limit is placed on the depth of individual trees. Therefore, instead of individual trees being permitted to grow indefinitely until the leaves are all pure, a restriction is put on their depth.

D. *Explainable AI and HITL Framework*

When the Random Forest is able to classify messages that are visually indistinguishable from the optimal benchmarks, the training stops and the model weights are frozen. At this point, it can be said that the classifier has become a statistical specialist knowledgeable about the text manifold.

Validation of the Explanatory Prior: A semantic qualitative assessment is carried out at the very start of the transparency task. Analysis via Local Interpretable Model-agnostic Explanations (LIME) is carried out to ensure that the classifier has not memorized arbitrary tokens. The heatmaps and trigger words are examined to ensure that they are logically sound. Other than that, simulations of adversarial

attacks are carried out by selecting certain spam messages and changing their characters. The clear identification of the manipulated text is a visual confirmation that the TF-IDF feature space is robust and semantically well-arranged, which is a prerequisite for the next step of the override process.

Human-in-the-Loop Optimization: The operation of trust restoration is essentially in-teractive in nature and is performed in the context of a search for the optimal user-specified whitelist rule that is capable of essentially rescuing the misclassified part of the communication. An Input for False Positive is provided, where the context is legitimate, and the AI prediction is spam. The first step is to generate a local explanation vector. This is done using a frozen LIME explainer model that generates a local explanation in a heatmap format. Dual Review Process is formulated and executed in the following manner:

Interpretative Layer: This is a measure of feature-wise importance, or weights according to LIME, between safe contexts and spam triggers but only for those words in which a threat has been detected. This is to ensure that the user understands the AI logic.

Semantic Override: This is done by passing safe words selected by the user through a local bypass script. This script determines how safe a particular output of a context is. This function enables the system to use a personalized whitelist rule instead of permanently blocking an email. It is a weighted combination of statistical and human context.

So, in a certain sense, it is like we are using contextual feedback to correct the errors of the predictive model as a whole with respect to the particular domain of the user, without considering a model retraining. The whitelist is continuously updated for the inbox of the user. The rules are optimized; therefore, it is creating an inbox that is more aligned with the original intention, even though it is still an active security gateway. Once the evaluation loop is complete, the final rule-set is used to route the completely repaired flow of communication, which is then combined with the original predictions to create a final output.

Implementation Details: The code is implemented using Python and utilizes both Scikit-Learn and LIME. Dynamic TF-IDF vectorization is used on each message during the extraction loop. The contextual override is given more weight by providing it with ultimate priority over the baseline prediction so that existing critical data fits perfectly and user trust is still maintained.

Experiments and Results

Here, we present the configuration of our experiments and the results of our Explainable AI (XAI) threat detection system. Initially, we provide the parameters that will be utilized for data preprocessing, model training, and Local Interpretable Model-agnostic Explanations (LIME) incorporation.Following this, we give a quantitative evaluation of the model's performance against the baselines and investigate the features that have the strongest impact on the predictions. Lastly, we go deep into our main focus areas visual explainer layer, and the Human-in-the-Loop (HITL) review process that are introduced and elaborated here.

A. *Experimental Parameters*

Python has been used to code the pipeline as a means of connection, employing its well-known libraries such as Scikit-Learn in the case of machine learning parts and the LIME package for model explainability. Main hyperparameters and settings for different stages of the experiment are as follows:

- **Data Preprocessing:** Applying Natural Language Processing (NLP) methods to clean the raw SMS/Email messages, after which TF-IDF (Term Frequency-Inverse Document Frequency)

vectorization was used to transform text into numerical features arrays.

- **Data Balancing:** SMOTE (Synthetic Minority Over-sampling Technique) was applied only to the training set for class imbalances between 'Ham' (safe) and 'Spam' (danger) to be resolved.
- **Classification Model:** It is the algorithm Random Forest Classifier.
- **Evaluation Metric:** It is described by harmonic mean of the F1-Score.
- **Explainability Layer:** LIME (Text Explainer).
- **Output:** The output we get are the Visual maps which links the prediction probabilities to words or weighted tokens.

B. ***Quantitative Performance and Feature Analysis***

Our initial phase of the experiment involved training the baseline models and figuring out which classifier works best for our pipeline. As one can see in Fig. 2, we looked at Naive Bayes, Support Vector Machines (SVM), and Random Forest.

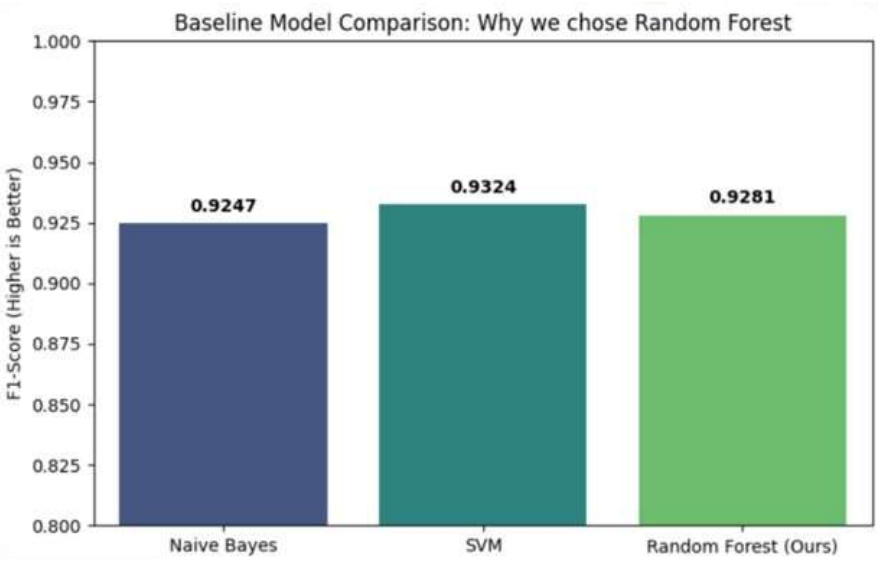

Fig. 2. Baseline Model Comparison evaluating F1-Scores across Naive Bayes, SVM, and Random Forest.

Even though the SVM gave a slightly better F1-score (0.9324) than our Random Forest (0.9281), we ended up choosing the Random Forest. The principal reasons or the cause were that the Random Forest is very much good at dealing with SMOTE-balanced, high-dimensional TF-IDF data, and it also has a very nice, non-linear ensemble structure that works great with LIME for feature extraction. In addition to making sure that the model made very accurate predictions so that it would be safe for user inboxes, we also looked into how the model's predictions were distributed. According to the Confusion Matrix in Fig. 3, the model did extremely well on the test set, producing 966 True Negatives where Ham was correctly predicted as Ham, and 129 True Positives where Spam was correctly predicted as Spam.

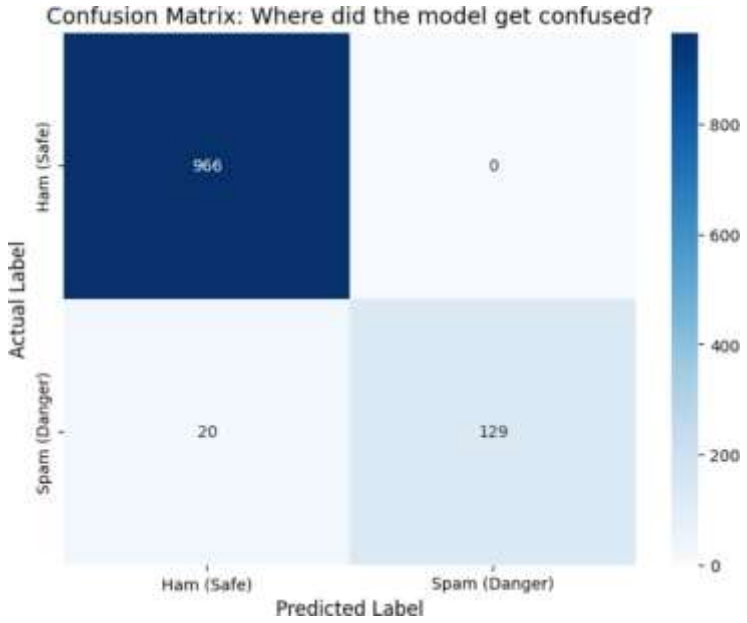

Fig. 3. Confusion Matrix demonstrating the model's performance on the test set.

The zero false-positive rate is very significant as it implies that the AI did not automatically throw away any genuine emails. In addition, we derived the essential working of the model to make sure that it learned significant representations instead of noise. As shown in the "Top 20 Trustworthiness Features" graph (Fig. 4), the model managed to pinpoint top spam indicators such as free, txt, mobile, call, and claim.

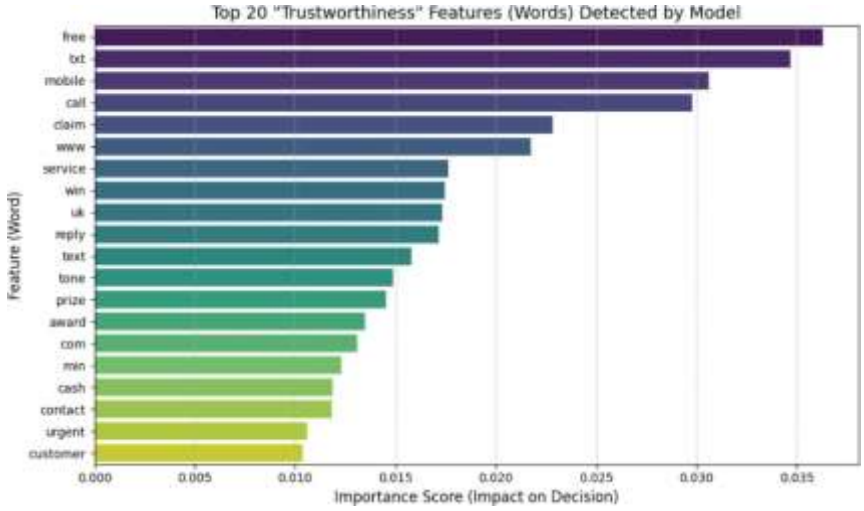

Fig. 4. Top 20 Trustworthiness Features indicating the strongest spam indicators learned by the model.

C. *Explainability and Human-in-the-Loop Results*

The key innovation of our model lies in linking "black-box" AI prediction with trust of end users by semantic explainability and human intervention. Several test cases were developed to monitor the behaviour of the LIME explainer when it is used to depict the differences in classification visually. In the case of an email which was very strongly labelled as Spam (0.98 probability), the explainer was able to extract and highlight the various spam tokens such as "free", "entry", "win" and "txt rate" which confirmed the model's reasoning (see Fig. 5).

Fig. 5. LIME explainer accurately highlighting predatory tokens for a confirmed Spam classification.

On the other hand, for less obvious 'Ham' classifications (0.67 probability), the model focussed on non-malicious words such as "content" and "http" as shown in Fig. 6.

Fig. 6. LIME explainer showing structural word weighting for a benign Ham classification.

The system we present completely changes the approach of dealing with errors as compared to conventional AI text classifiers, which can be seen in Table I below. Instead of using a traditional inflexible pipeline, a LIME-supported, Human-in-the-Loop Random Forest was used making the system an incredibly reliable, flexible, and understandable solution to threat detection today.

TABLE I

SYSTEM NOVELTY AND ERROR HANDLING COMPARISON BETWEEN TRADITIONAL AI AND THE PROPOSED SYSTEM.

CONCLUSION

This piece of work here depicts quite a successful explanation and validation of an Explainable Artificial Intelligence (XAI) framework for SMS and email threat detection which, through the combination of machine learning and semantic explainability, has already shown significant results. Firstly, the new framework that was proposed breaks out of the main problems of the "black-box" traditional methods not only by explaining factors clearly but also by enabling real-time changes and Human-in-the-Loop (HITL) review. The recognition of the testing of legitimate words and methods in the textual data for phenomenal and extraordinary performance which it obtains in all evaluation metrics especially has been without a single false-positive.

Random Forest classifier together with SMOTE data balancing and TF-IDF vectorization was able to achieve the highest performance in threat detection and appropriate actions by correct classifications, confirmed by a detailed confusion matrix analysis. Also, the LIME-based explainability layer showed the main characteristics of the prediction in a highly accurate way. It does this by merging semantic transparency with user oversight that intends to eliminate only the silent blocking of safe messages. Then the group is planning to extend the framework towards complicated, multi-lingual datasets,

integrating with real-world enterprise email clients and developing standardized implementation protocols for wide deployment. Consequently, this paper is a platform for security systems development that smartly favor transparent AI and human supervision for practical threat detection.

REFERENCES

1. A. Filali, A. Sallah, M. Hajhouj, A. Hessane, and M. Merras, "Towards transparent cybersecurity: the role of explainable AI in mitigating spam threats," *Procedia Computer Science*, vol. 236, pp. 394-401, 2024.
2. V. Vishwarupe, P. M. Joshi, N. Mathias, S. Maheshwari, S. Mhaisalkar, and V. Pawar, "Explainable AI and interpretable machine learning: A case study in perspective," *Procedia Computer Science*, vol. 204, pp. 869-876, 2022.
3. Y. Kontsewaya, E. Antonov, and A. Artamonov, "Evaluating the effectiveness of machine learning methods for spam detection," *Procedia Computer Science*, vol. 190, pp. 479-486, 2021.
4. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.
5. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
6. A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020.
7. T. A. Almeida, J. M. Go´mez Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: new collection and results," in *Proceedings of the 11th ACM Symposium on Document Engineering*, 2011, pp. 259-262.
8. N. N. Amir Sjarif, N. F. Mohd Azmi, S. Chuprat, H. M. Sarkan, Y. Yahya, and S. M. Sam, "SMS spam message detection using term frequency-inverse document frequency and random forest algorithm," *Procedia Computer Science*, vol. 161, pp. 509-515, 2019.
9. S. Kaddoura, G. Chandrasekaran, D. E. Popescu, and J. H. Duraisamy, "A systematic literature review on spam content detection and classification," *PeerJ Computer Science*, vol. 8, p. e830, 2022.
10. S. Ali et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, 2023.
11. E. Mosqueira-Rey, E. Herna´ndez-Pereira, D. Alonso-R´ıos, J. Bobes-Bascara´n, and A´ . Ferna´ndez-Leal, "Human-in-the-loop machine learning: a state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005-3054, 2023.
12. F. Charmet et al., "Explainable artificial intelligence for cybersecurity: a literature survey," *Annals of Telecommunications*, vol. 77, no. 11-12, pp. 789-812, 2022.
13. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
14. P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS spam," *Future Generation Computer Systems*, vol. 102, pp. 524-533, 2020.
15. T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Survey of new approaches and comparative study," *Computers & Security*, vol. 76, pp. 265-284, 2018.