Iqra Journal of Engineering & Computing
Vol. 02, No. 1, (March 2026) 53 – 73
DOI 10.21621/ijec.20260101.04

Performance Evaluation of Machine Learning for Cross-Lingual Ham and Spam Detection

Kashif Iqbal[1*], Sikandar Ali[2], Mustafa Haider Ali[3], Muhammad Hammad[3], Faraz Ali [3], Raheema Agha[4]

ABSTRACT

This paper presents a multilingual spam email classification framework capable of processing both German and English texts using classical machine learning models combined with a translation-enhanced pipeline. The system integrates TF-IDF vectorization, Logistic Regression, Multinomial Naive Bayes, and a probability-calibrated Linear Support Vector Classifier (SVC). A key contribution is the incorporation of an automatic German-to-English translation layer using MarianMT, allowing cross-lingual evaluation and robustness analysis. Multiple models are trained on a manually curated dataset of 3,790 emails, achieving up to 98.55% accuracy and 0.9860 AUC-ROC on the evaluated dataset. A Streamlit-based application is implemented for real-time inference, supporting auto-detection of input language, dual-model evaluation for German emails, and calibrated probability scores. Experimental results demonstrate that Logistic Regression provides the most consistent overall performance, while Calibrated Linear SVC delivers the highest AUC-ROC and stable decision boundaries. The system represents a practical, expandable multilingual spam detection pipeline suitable for lightweight deployment scenarios.

Keywords: Spam classification, TF-IDF, Logistic Regression, Na¨ıve Bayes, Calibrated Linear SVC, Calibrated Classifier CV, machine translation, MarianMT, multilingual NLP, Streamlit, Deep Learning, BERT.

Author's Affiliation:
Institution(s) Name: [1]Department of Computer Science & A.I, BIC, Karachi,
[2]Department of Law, BIC, Karachi, Pakistan.
[3]Department of Computer Science, SZABIST, Karachi,
[4]Department of Management Science, BIC, Karachi,
Country: Pakistan
Corresponding Author's Email: [1] kashif.iqbal@bic.edu.pk

Published by ORIC, Iqra University, Main Campus, Karachi-Pakistan.

3105-4528 (Online), 3105-451X (Print) © 2026.

1. Introduction

Email remains one of the most widely used communication mediums for both personal and professional correspondence. Alongside its benefits, unsolicited and malicious spam emails continue to pose significant challenges, including financial fraud, phishing attacks, malware distribution, and privacy violations. These unwanted emails sent to a user's email account. These unsolicited emails are called spam Almost 57% of the world's email traffic [1]. Prior studies have shown that spam constitutes a substantial portion of global email traffic, motivating continuous research into reliable and efficient spam detection systems [3] [4].

Early spam filtering techniques relied on rule-based systems and handcrafted heuristics, which proved difficult to maintain and adapt to evolving spam patterns. Consequently, machine learning approaches such as MNB, SVM, and LR became dominant due to their ability to learn discriminative patterns directly from textual data [5] [6] [2]. When combined with sparse lexical features such as TF-IDF, these classical models have demonstrated strong performance while remaining computationally efficient and interpretable [6] [7].

Despite extensive research on monolingual spam classification, multilingual and cross-lingual spam detection remains comparatively underexplored. Most existing systems are designed for English-only datasets, limiting their applicability in multilingual regions where users frequently receive emails in multiple languages [8]. Training and maintaining separate classifiers for each language introduces scalability and maintenance challenges. Recent studies indicate that machine translation can serve as an effective bridge for cross-lingual text classification, allowing models trained in one language to generalize across others [8] [9]. Translation-assisted pipelines reduce the need for language-specific classifiers while allowing comparative evaluation across languages.

However, much of the existing multilingual text classification research focuses on deep neural architectures, which often require large datasets, extensive computational resources, and complex deployment pipelines. In contrast, practical deployment environments, such as real-time email filtering systems, benefit from lightweight, interpretable, and easily deployable solutions [19]. Classical machine learning models remain attractive in such contexts due to their low latency, transparency, and

robustness in limited datasets. Motivated by these gaps, this work proposes a cross-lingual spam classification framework that combines classical machine learning models with a translation-enhanced pipeline. The system supports both German and English emails, applies TF-IDF-based feature extraction, and evaluates Logistic Regression, Multinomial Naıve Bayes, and probability-calibrated Linear SVC models. A real time Streamlit application is developed to demonstrate practical usability, enabling language detection, dual-model evaluation for German emails, and interpretable probability-based predictions.

The primary objectives of this work are:

- Design a multilingual email spam detection pipeline.
- Comparison of LR, NB, and Calibrated Linear SVC for both German and English datasets.
- Evaluate the effect of machine translation on behavior of the classifier.
- Developing a real-time app supports language detection and probability-based predictions.

Spam email detection is a widely studied problem; however, most existing solutions focus primarily on English-language datasets or rely on computationally expensive deep learning models. During experimentation with classical spam classifiers, it became evident that real-world email communication is often multilingual, while many traditional pipelines assume a single-language setting. Another practical limitation observed was that multilingual spam detection typically requires training and maintaining separate models for each language, increasing system complexity. At the same time, there is limited applied research that analyzes whether classical machine learning models can remain effective in a cross-lingual setting when combined with machine translation. Additionally, many studies remain confined to offline evaluation and do not demonstrate how multilingual spam classifiers behave in an interactive, real time environment. These observations motivated the development of a lightweight, interpretable, and deployable multilingual spam classification framework that supports cross-lingual analysis while remaining practical for real-world use.

Motivated by the need for a practical and lightweight multilingual spam detection system, this project makes the following contributions:

- C1. Design and implementation of a multilingual spam classification pipeline capable of processing both German and English email content using classical machine learning models and text representation-based TF-IDF.
- C2. A systematic comparison of Logistic Regression, Multinomial Naive Bayes, and Calibrated Linear Support Vector Classifier to analyze their performance, robustness, and error behavior across German and English datasets.
- C3. Integration of a machine translation–based approach using MarianMT to enable cross-lingual spam classification, allowing German emails to be evaluated using English-trained models without retraining multilingual classifiers.
- C4. Development of a real-time, user-facing application that combines language detection, translation, and model inference, enabling interactive experimentation, model comparison, and probability-based interpretation of spam predictions.

2. Related Work

2.1 Traditional and Ensembles Machine Learning

Although deep neural networks have been introduced, classical machine learning models are highly applicable, since they are simpler to understand and calculate. Ensemble techniques are still being used to optimize these models by researchers. For instance, Adnan et al. [19] had decided to use model's logistic regression, decision tree, k-nearest neighbors (KNN), Gaussian naïve Bayes, and Adaboost. Moreover, they have chosen two distinct datasets from spam emails. Iqbal et al [20], Yu et al. [24] and P. Ghogare et al. [35] while conducting study on spam classification appointed Naïve Bayes, Support Vector Machines (SVM), Random Forest, Adaboost and XGBoost as their algorithms. While using these, the particular datasets selected by them is CodEAlltagXL email corpus. Tian et al [12] suggested a stacking ensemble methodology to enhance the accuracy of detecting spam email in cybersecurity. Their contribution underlines the fact that the combination of base classifiers may result in the development of strong decision boundaries, a concept that is consistent with the combination of various probabilistic models (e.g., Naive Bayes and Logistic Regression) to guarantee the consistency of performance.

In addition, extensive searches, one of which was conducted by Tusher et al [13], Pantel et al. [22] and Wang et al. [30] selected Naïve Bayes algorithm while using Spam Cop as a dataset, Janez-Martino et al. [27] appointed TF-IDF combined with SVM. The very first multi-class dataset SPMEC-11K applied during this study. Krishnamoorthy et at. [32] selected several categories Logistic Regression (LR), Convolutional Neural Networks (CNN), Random Forests (RF), Recurrent Neural Networks (RNN), Long Short-Term Memory. (LSTM), and suggested Deep Neural Networks (DNN). Iswanto et al. [31] makes use of various algorithm to find out which one gives better accuracy such as RF, KNN and Naïve Bayes. The study used datasets derived from GitHub. Mujtaba et al. [28] utilized PU, SpamAssasin, Spambase, TREC, Enron, CCERT, LingSpam, and Phishing Corpus as Datasets. The algorithms used were ML, DL and GA, note that despite the fact that the detection techniques have been fine-tuned to the best of their ability, challenges such as computational overhead, the shifting spammer evasion techniques, and data set disproportion remain significant research challenges.

2.2 Transformer-Based and Deep Learning

Transformer models and deep learning architectures have been widely used as the paradigm shift in spam classification to natural language processing (NLP). The effectiveness of the fine-tuned transformer models with the attention mechanisms in email spam classification was demonstrated by Shah [14] where they are most effective in the context of the complex semantic context. Continuing on the uses of transformers, Khan et al. [18] have preferred BERT (Bidirectional Encoder Representations from Transformers) and LSTM. In order to carry out this particular research on spam Detection. The datasets that they have considered are PU, Lingspam, and Enron while using accuracy, recall, precision and F1 score as performance measures. Jamal et al [15].

Similarly, Rashed and Ozcan [16] introduced a novel dual-layer deep learning model, which combines LSTM-based networks with (GPT-2 and XLM-RoBERTa) to detect phishing and spam email messages, in particular, those that contain a false attachment. Kameswari et al. [21] used numerous deep learning techniques for multilingual spam classification such as RNN, LSTM, Bi-LSTM and GRU. The datasets consisted of four languages English, Hindi, French and German. Haque et al [23] have picked ANN, CNN, LSTM, GRU, Bi-LSTM model for spam classification. The datasets utilized during this were spam base and Lingspam. Malhotra et al. [33] importantly considers Natural language processing (NLP), using dense classifier sequential neural network, LSTM and Bi-LSTM as well. Kumar et al. [33] utilize MLP, Naïve Bayes, and Random Forest to acquire Better accuracy. Zhang et al. [34] make use of a new multilingual spam detection dataset comprising over 30,000 samples. They chose convolutional neural networks (CNNs), and Linear support Vector machines (SVMS) for image-based spam filtering. Attique et al. [35] introduced big data-driven insights into email classification. The Learning algorithms used in this study were DT, NB, and ANN. Despite the fact that these models have achieved the state of the art accuracy, they are computationally expensive, and this is a big challenge to deploy to lightweight devices as well as in real-time.

2.3 Multimodal and Advanced Architectures

Contemporary spam can also be multimedia-based in addition to the textual content to evade text-based filters. Gahara et al [17]. This was taken care of (2025) by suggesting a hierarchical multimodal robust spam detector that uses LLMs and CNNs. Roumeliotis et al. [25] applied GPT-4 Large language model (LLM), along with BERT and ROBERTa Natural Language Processing (NLP) methods. Rojas et al. [29] decided on using Large Language models (LLMs) of spam classification. They also analyse the performance of both open-source (Flan-T5) and proprietary LLMs (ChatGPT, GPT-4) while considering Spam Assassin as their main Dataset.

3. Methodology

3.1 Dataset Description

The dataset utilized in this study the codeAlltag German corpus [10] [11], a dataset comprising of 3,790 email messages, each manually annotated for supervised spam classification. Out of these, 2,997 emails are labeled as legitimate (ham), while 793 emails are labeled as spam, resulting in a moderately imbalanced dataset that reflects real-world email distributions. The emails include a mixture of personal correspondence, transactional messages, phishing attempts, promotional spam, lottery scams, and account security warnings. Each dataset instance contains two primary attributes: (i) the raw email text, representing the unstructured content of the message, and (ii) a binary class label, where 0 denotes a legitimate (ham) email and 1 denotes a spam email. The dataset includes both German language emails and their English counterparts, enabling multilingual and cross-lingual experimentation.

3.2 Problem Formulation

Spam email detection is formulated as a binary supervised text classification problem. Given an email message represented as a text sequence x, the objective is to predict its corresponding class label y, where

$$y \in \{0, 1\} \quad (1)$$

with $y = 0$ indicating a legitimate (ham) email and $y = 1$ indicating a spam email. After preprocessing, each email is transformed into a high-dimensional numerical feature vector using the Term Frequency–Inverse Document Frequency

Table 1: Algorithm and dataset Description

Ref	Year	Algorithm	Datasets	Performance Metrics
[18]	2022	BERT, LSTM	PU, LingSpam, Enron	Accuracy, recall, precision and f1-score
[19]	2024	LR, DT, KNN, NB,AB	Two distinct datasets selected from spam emails	recall, precision and F1-score
[20]	2021	NB, RF, SVM and DL	CodEAIItagXL email corpus	Accuracy, precision, recall and f1- score
[20]	2025	NB, RF, SVM, AB, XGB	CodEAIItagXL	Accuracy, precision, recall and f1- score
[12]	2024	RNN, LSTM, Bi- LSTM, GRU	Datasets were consisting of spam and ham messages	Accuracy, Recall
[12]	1998	NB	SpamCop	Accuracy
[23]	2025	ANN, CNN, LSTM, GRU, Bi-LSTM	Spam base and Lingspam	Accuracy, precision, recall, f1-score
[24]	2028	NB, SVM, NN	spam filtering corpora Spam Assassin and Babletext	Accuracy, precision, recall
[25]	2024	LLM, BERT, RoBERTa, NLP	Ham/Spam Emails from Kaggle.	Accuracy, precision, recall, f1-score
[26]	2024	LR, NB, RF, ANNs	2007 TREC Public Spam Corpus and Enron-Spam.	F1-score, recall, precision and accuracy
[32]	2024	LR, CNN, RF, RNN, LSTM, DNN	7 Enron datasets	Accuracy
[27]	2020	SVM, NB	SPMEC-11K	F1 - score

[28]	2017	SVM, DT, GA, ANN, NB, RF	PU, SpamAssasin, Spambase, TREC, Enron, CCERT	Pr, Rec, FS, specificity, AUC, AA, ER
[34]	2023	SVM, CNN	Multilingual dataset comprising over 30,000 samples	Accuracy, precision, recall and f1- score
[29]	2024	LLMs	SpamAssasin	Accuracy,BA, precision, recall and f1-score
[35]	2023	DT, NB, ANN	Spambase, Enron and Spam7	Accuracy, precision, recall and f1-score
[30]	2025	NB	TREC and Enron	Accuracy, precision, recall and f1-score
[31]	2024	KNN, NB	Datasets were derived from GitHub	Accuracy, precision, recall and f1-score
[35]	2023	SVM, NB, RF	Enron and SpamAssasin	Accuracy, precision, recall, specificity and FS
[26]	2021	NB, RF, MLP	Enron 1	Accuracy, precision, recall, ROC and FS
[33]	2022	LSTM, BI-LSTM	Spam Emails	Accuracy, recall, and FS

(TF–IDF) representation. Let $x \in \mathrm{R}^d$ denote the TF–IDF feature vector corresponding to an email, where d is the size of the learned vocabulary. A classifier $f(\cdot)$ is trained to approximate the mapping:

$$f(x) \rightarrow y \quad (2)$$

such that the predicted label y accurately reflects whether the input email is spam or legitimate.

For German-language emails, the problem is extended to a cross lingual classification setting. A machine translation function $T(\cdot)$ is introduced to map German text into English:

$$x_{en} = T(x_{de}) \quad (3)$$

where x_{de} denotes the original German email and x_{en} represents its English translation.

3.3 Modeling Techniques

Logistic Regression: Logistic Regression is employed as a baseline discriminative classifier due to its simplicity, interpretability, and strong performance on high-dimensional sparse text representations. In the context of spam classification, Logistic Regression models the conditional probability of an email being spam

given its TF–IDF feature representation. Formally, given an input feature vector $x \in R^d$, Logistic Regression estimates the probability of the spam class as:

Multinomial Na¨ıve Bayes: Multinomial Na¨ıve Bayes is a probabilistic generative model widely used in text classification tasks due to its computational efficiency and robustness on small datasets. The probability of a document $x = (x_1, x_2, \ldots, x_d)$ belonging to class $c \in \{0, 1\}$ is computed as:

Calibrated Linear Support Vector Classifier: In this study, a Linear Support Vector Classifier (Linear SVC) is utilized due to its scalability and
effectiveness on large, sparse TF–IDF feature spaces. To address the lack of probabilistic outputs in standard SVMs, the Linear SVC is wrapped using CalibratedClassifierCV, enabling probability estima- tion via Platt scaling. The decision function is defined as:

$$f(x) = w^T x + b \qquad (6)$$

Classification is performed based on the sign of $f(x)$, while probability calibration maps the decision scores to posterior probabilities.

Fine-tuned for sequence classification. The model tokenizes both English and German inputs using its native sub-word tokenizer and passes the pooled output through a linear classification head. The inclusion of this deep learning baseline allows for a direct comparison between computationally expensive transformer architectures and our pro- posed lightweight, TF-IDF-based classical machine learning pipeline.

3.4 Machine Translation Pipeline

A key component of this architecture is the German-to- English translation using MarianMT. The pipeline is visually detailed in Figure 1 and follows these steps:

1) Detect language of input email.
2) If German:

- Classify German text using German models.
- Translate email to English.
- Classify translated version using English models.

Display both outputs for cross-lingual consistency.
If the detected language is English, the system bypasses the translation step and directly applies the English TF-IDF vectorizer and user selected classification model. The predicted class (spam or ham) along with

calibrated probability scores is displayed to the user. Figures 3, 4, and 5 illustrate the fully functional application interface in action.

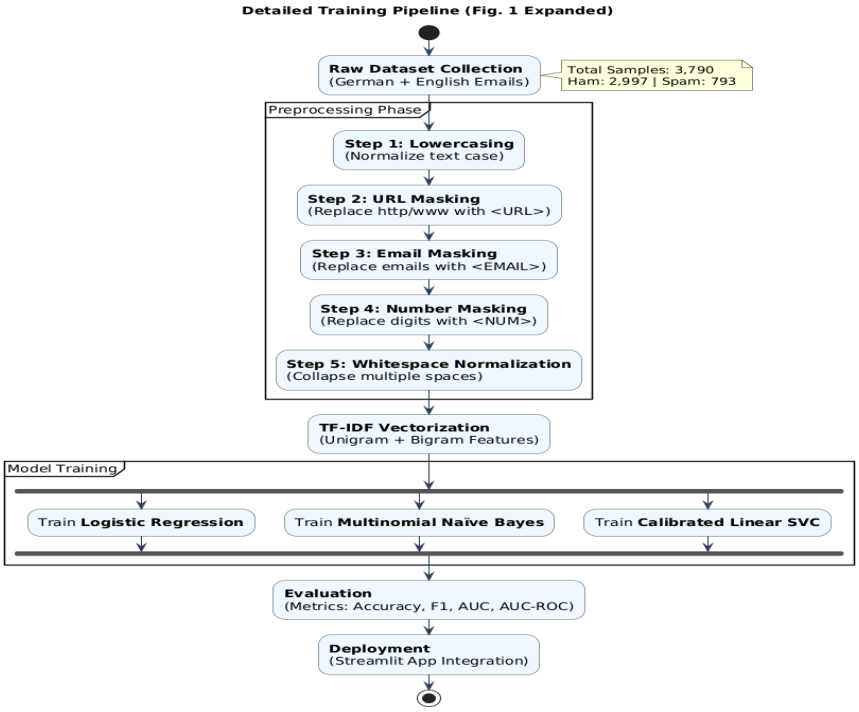

Figure 1. Training and Inference Pipeline of the Proposed Model.

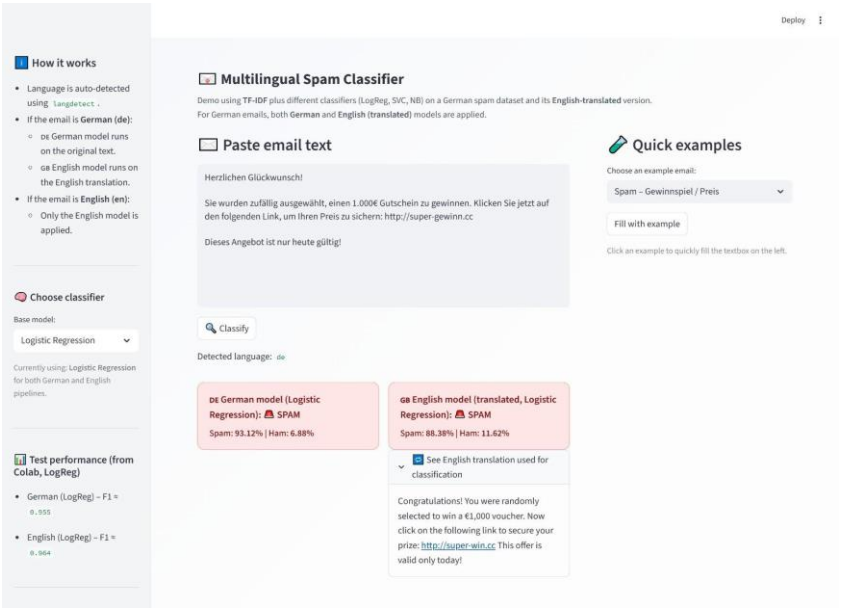

Figure 2. Streamlit application interface processing an input email

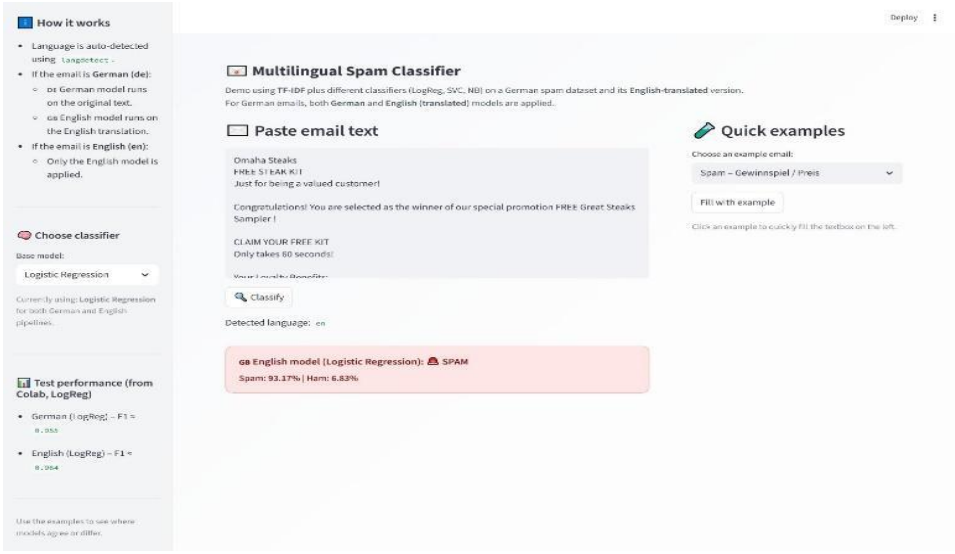

Figure 3. Streamlit application demonstrating language detection and model evaluation

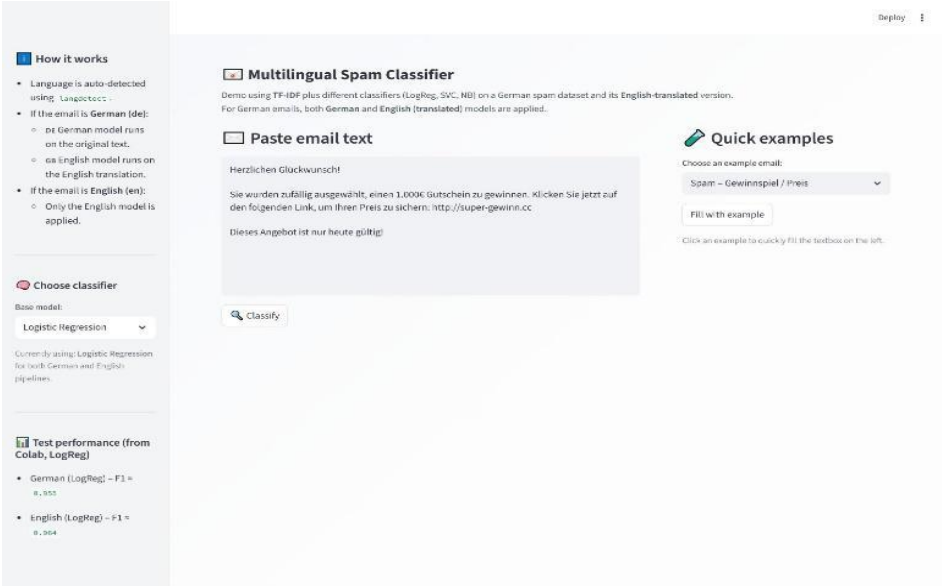

Figure 4 Stream lit application after detecting language.

.

4. Evaluation Metrics

To comprehensively assess the performance and reliability of the classification models, several standard statistical metrics are utilized.

4.1 Precision

Precision quantifies the exactness of the model's positive predictions. It indicates the proportion of instances correctly classified as spam out of all the instances the model flagged as spam. A high precision rate is critical in email filtering to minimize false positives (legitimate emails incorrectly marked as spam).

$$Precision = \frac{TP}{TP + FP}$$

where 'TP' represents True Positives and 'FP' represents False Positives.

4.2 Recall (Sensitivity)

Recall, also referred to as sensitivity or the true positive rate, measures the model's capacity to successfully detect all actual positive cases. It emphasizes the ratio of correctly identified spam emails to the total number of actual spam emails present in the dataset.

$$Recall = \frac{TP}{TP + FN}$$

where 'FN' represents False Negatives.

4.3 Specificity

Specificity evaluates the model's proficiency in correctly identifying negative cases (legitimate, non-spam emails). It determines the percentage of actual ham emails that the model successfully rejected from the spam classification, reflecting the system's ability to recognize safe content.

$$Specificity = \frac{TN}{TN + FP}$$

where 'TN' represents True Negatives.

4.4 F1-Score

The F1-score serves as a harmonic mean of precision and recall. It provides a single, composite metric that effectively balances the trade-off between false positives and false negatives, making it highly valuable when evaluating moderately imbalanced datasets.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.5 AUC and ROC

The Receiver Operating Characteristic (ROC) curve is a graphical plot illustrating the diagnostic ability of a binary classifier as its discrimination threshold varies. It plots the True Positive Rate (Recall) against the False Positive Rate (1 - Specificity). The Area Under the Curve (AUC) summarizes the ROC curve into a single scalar value representing the model's overall degree of separability. An AUC value closer to 1.0 implies an outstanding model capable of perfectly distinguishing between classes, while a value near 0.5 suggests performance equivalent to random guessing.

4.6 NPV

Negative Predictive Value (NPV) is described as the ratio of truly negative results to all negative predictions made by the model. NPV acts as an important criterion for assessing classifier accuracy, along with Precision, Recall, and F1 score. It evaluates the ability of a model to predict negative cases, especially highlighting the likelihood that a negative prediction is truly negative. This metric holds significance in cases where data sets are unbalanced or diagnoses are concerned, and avoiding false negatives becomes important.

$$NPV = \frac{TrueNegatives}{TrueNegatives + FalseNegatives}$$

Table 2: Precision, Recall and NPV

Model	Precision	Recall	NPV
LogReg	1.0000	0.9308	0.9819
Calib. SVC	0.9932	0.9245	0.9803
BERT (Baseline)	0.9722	0.8805	0.9691
MNB	1.0000	0.7861	0.9463

Table 3: Accuracy, F1 score and AUC

Model	Accuracy	F1 (Spam)	AUC
LogReg	0.9855	0.9642	0.9821
Calib. SVC	0.9828	0.9577	0.9860
BERT (Baseline)	0.9697	0.9241	0.9741
MNB	0.9551	0.8803	0.9704

Table 4: Accuracy, F1 score and AUC

Model	Accuracy	F1 (Spam)	AUC
LogReg	0.9815	0.9548	0.9820
Calib. SVC	0.9789	0.9480	0.9842
BERT (Baseline)	0.9683	0.9205	0.9758
MNB	0.9696	0.9220	0.9657

Table 5: Precision, Recall and NPV

Model	Precision	Recall	NPV
LogReg	0.9801	0.9308	0.9819
Calib. SVC	0.9799	0.9182	0.9786
BERT (Baseline)	0.9720	0.8742	0.9675
MNB	1.0000	0.8553	0.9630

Table 6: Accuracy, F1 score and AUC for German and English

Language	Model	Accuracy	F1 (Spam)	AUC
English	LogReg	0.9855	0.9642	0.9821
German	LogReg	0.9815	0.9548	0.9820

Table 7: Accuracy, F1 score and AUC for German and English

Model	Accuracy (EN-DE)	F1
LogReg	0.0040	0.0094
Calib. SVC	0.0039	0.0097
BERT (Baseline)	0.0014	0.0036
MNB	0.0145	0.0417

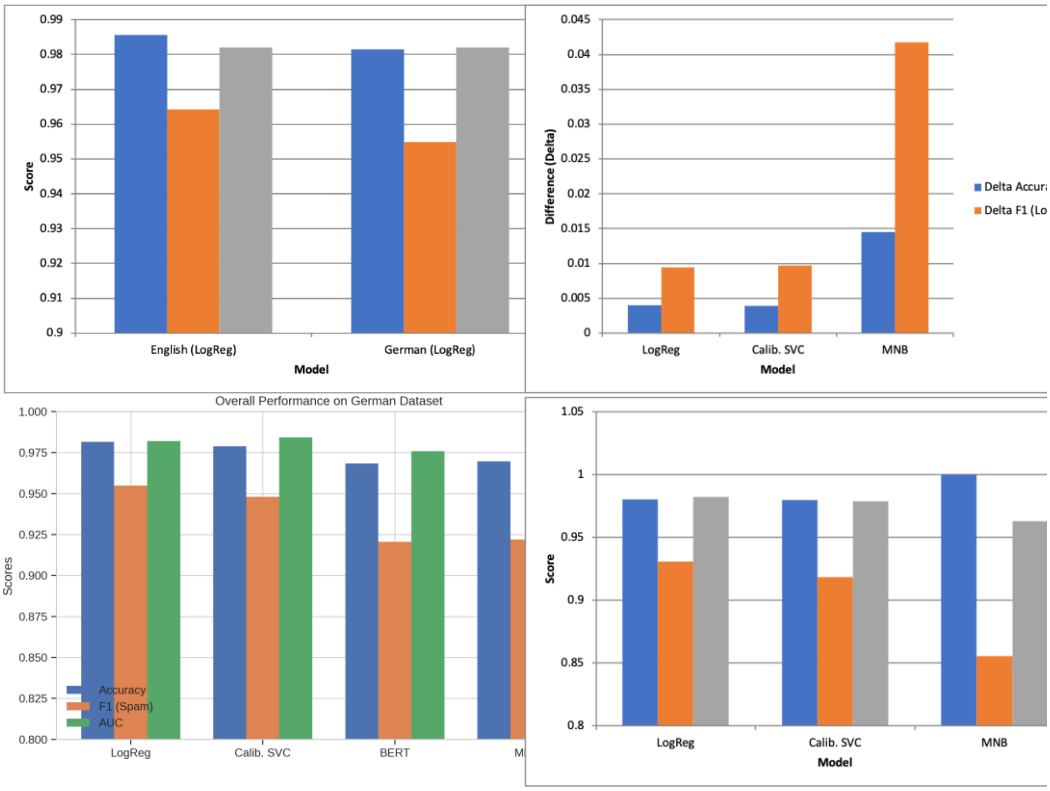

Figure 5. Model selection and Performance

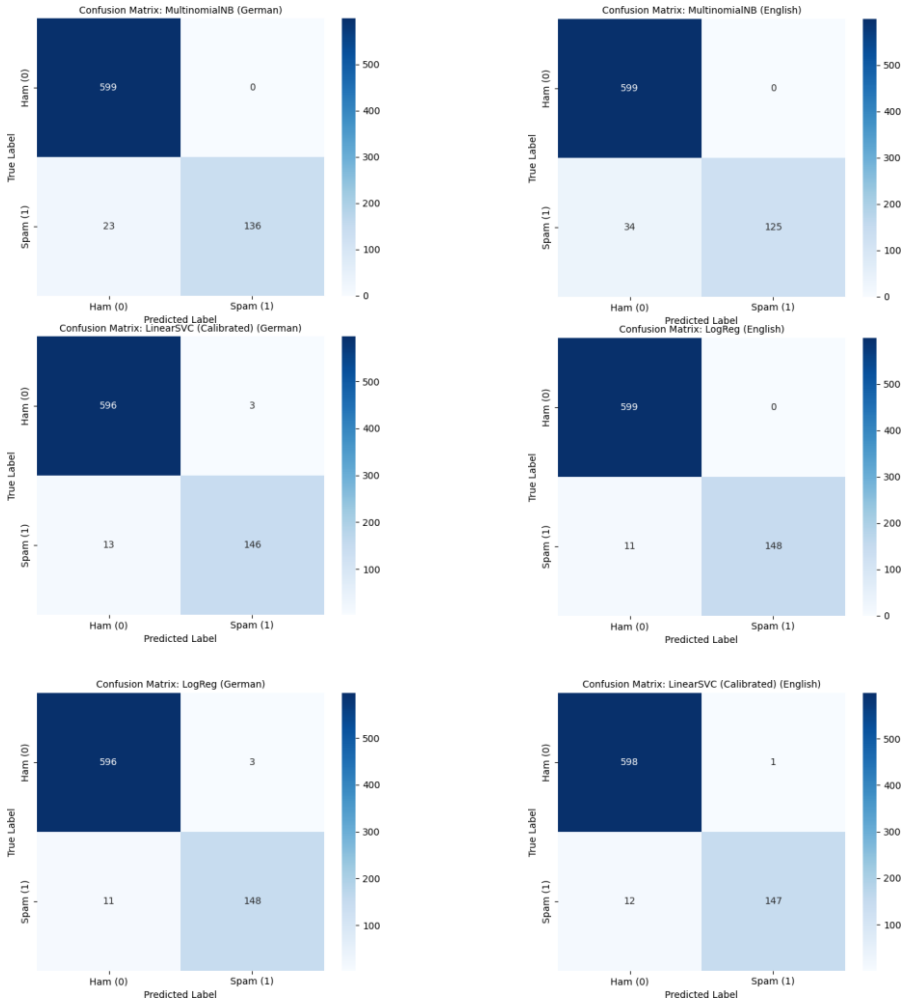

Figure 6: Confusion Matrix of Spam Dataset

5. Results and Analysis

The performance of the classification models on the English test set was evaluated using Accuracy, F1-score (for the Spam class), and Area Under the Curve (AUC). As shown in Table II, Logistic Regression (LogReg) demonstrated the most robust performance across all metrics, achieving an accuracy of 0.9855 and the highest F1-score of 0.9642. The Calibrated Linear SVC followed closely with an accuracy of 0.9828 and the highest AUC of 0.9860. Figure 6 visually compares these overall metrics, highlighting the strong discriminative power of the linear models over Multinomial Naïve Bayes.

We also analyzed specific spam detection metrics including Precision, Recall, and Negative Predictive Value (NPV), as presented in Table III. Logistic Regression achieved perfect precision (1.0000) alongside a high recall of 0.9308. This exceptional precision signifies that no legitimate emails were falsely categorized as spam, which is a vital requirement for user trust in email filtering systems. Figure 7 illustrates this trade-off between Precision and Recall for the English models.

Confusion matrices for both languages are shown to highlight exact true/false positive distributions. Figure 8 displays the behavior of the models natively on the German text, while Figure 9 shows the results for the English text. These matrices explicitly visualize how the models manage class separation, further confirming the low false-positive rates of Logistic Regression.

The evaluation was repeated for the German dataset to assess the models' native language capabilities. As detailed in Table IV, Logistic Regression again outperformed the other classifiers with an accuracy of 0.9815 and an F1-score of 0.9548. Figure 10 summarizes the overall performance metrics for the German models.

Table V provides the granular spam detection metrics for the German dataset. The Calibrated SVC showed a slight drop in recall (0.9182) compared to Logistic Regression (0.9308). Figure 11 displays these trends, reconfirming that the models behave consistently regardless of whether the text is naturally German or computationally translated into English.

Table 6 consolidates the best-performing models across both languages. Logistic Regression emerged as the superior model in both English and German environments. Finally, we evaluated the cross-lingual stability of the models. Stability ('Delta') is defined as the absolute difference in performance metrics between the English and German datasets, calculated as:

$$\Delta Metric = |Metric_{EN} - Metric_{DE}|$$

As shown in Table VII, Logistic Regression and Calibrated SVC demonstrated exceptional stability with 'Delta' Accuracy values of roughly 0.0040. The minimal difference in the 'Delta' metrics for these linear models confirms that the translation layer does not introduce significant informational loss, allowing the English-trained model to evaluate translated German text with near-native accuracy. In contrast, MNB showed higher instability ('Delta F1 = 0.0417).

6. Conclusion

This study evaluated a multilingual spam classification system using classical machine learning techniques and compared them against a deep learning baseline (BERT) comparative analysis revealed that Logistic Regression delivered the most optimal and balanced classification performance, unexpectedly outperforming the computationally heavy BERT architecture on this specific dataset. The Calibrated Linear SVC proved highly effective in establishing precise decision boundaries. The integration of an automatic German- to-English translation layer enabled a robust cross-lingual evaluation, validating the system's consistency across language barriers. The deployment of this framework via an interactive application further demonstrated that practical, real-time multi- lingual spam detection can be achieved efficiently without the computational overhead typically demanded by deep neural networks.

7. Future Work

While the proposed multilingual spam classification framework demonstrates highly promising results using classical machine learning algorithms, several avenues exist for further enhancement. Primarily, the scope of the research can be expanded by increasing the volume and linguistic diversity of the dataset, incorporating languages beyond English and German. This expansion would facilitate a more rigorous evaluation of the models' cross-lingual robustness and adaptability. Additionally, integration of advanced preprocessing mechanisms—

such as semantic normalization, context-sensitive token filtering, and language-specific stopword handling—could significantly refine classification accuracy, particularly for textually ambiguous emails. Another promising direction involves developing and benchmarking deep learning-based multilingual architectures against the classical models evaluated in this study. For real-time deployment scenarios, future iterations could focus on optimizing translation latency and introducing incremental learning capabilities, allowing the system to dynamically adapt to evolving spam patterns based on continuous user feedback. Ultimately, transitioning this framework into a fully automated, production-scale spam filtering environment will provide deeper insights into its long-term scalability and operational usability

Funding: No external funding was declared for this manuscript.

Conflict of Interest: The author declares no known conflict of interest.

Data Availability: This framework-based study does not rely on private organizational data. The scoring examples are demonstrative and are included only to explain the proposed model. Future empirical validation will require authorized case-study datasets or controlled benchmark applications.

Ethics Statement: This study does not involve human participants, patient records, animal subjects, or private organizational data. Any future case-study testing must be performed only on owned systems or systems with written authorization.

Author Contributions: The author is responsible for conceptualization, framework design, manuscript preparation, and final review.

Acknowledgment: The author acknowledges the role of all authors

References

[1] W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang, "A support vector machine based naive Bayes algorithm for spam filtering," in *Proc. IEEE 35th Int. Performance Computing and Communications Conf. (IPCCC)*, Las Vegas, NV, USA, Dec. 2016, pp. 1–8.

[2] S. Chakraborty and B. Mondal, "Spam mail filtering technique using different decision tree classifiers through data mining approach—A comparative performance analysis," *International Journal of Computer Applications*, vol. 47, no. 16, pp. 1–7, 2012.

[3] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[4] T. Joachims, "Text categorization with support vector machines," in *Proc. European Conf. Machine Learning (ECML)*, Chemnitz, Germany, 1998, pp. 137–142.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint* arXiv:1301.3781, 2013.

[6] J. Tiedemann and O. De Gibert, "The OPUS-MT dashboard—A toolkit for a systematic evaluation of open machine translation models," in *Proc. 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Toronto, ON, Canada, Jul. 2023, pp. 315–327.

[7] *Scikit-learn Documentation*. [Online]. Available: Scikit-learn Documentation

[8] *MarianMT Framework*. [Online]. Available: MarianMT Framework

[9] *Streamlit Documentation*. [Online]. Available: Streamlit Documentation

[10] *CodE Alltag German Corpus*. [Online]. Available: CodE Alltag German Corpus GitHub

[11] E. Eder, U. Krieg-Holz, and U. Hahn, "CodE Alltag 2.0—A pseudonymized German-language email corpus," in *Proc. 12th Language Resources and Evaluation Conf. (LREC)*, Marseille, France, May 2020, pp. 4466–4477.

[12] Y. Tian, X. Dai, Z. Li, H. Guo, and X. Mao, "Improving the accuracy of cybersecurity spam email detection using ensemble techniques: A stacking approach machine learning for spam email detection," *PLoS One*, vol. 20, no. 9, p. e0331574, 2025.

[13] E. H. Tusher, M. A. Ismail, M. A. Rahman, A. H. Alenezi, and M. Uddin, "Email spam: A comprehensive review of optimized detection methods, challenges, and open research problems," *IEEE Access*, vol. 12, pp. 143627–143657, 2024.

[14] S. S. Shah, "Email spam detection: Leveraging fine-tuned transformer models with attention mechanism," Ph.D. dissertation, National College of Ireland, Dublin, Ireland, 2025.

[15] S. Jamal, H. Wimmer, and I. H. Sarker, "A large language model based on an improved transformer model to detect phishing, spam and ham emails," *Security and Privacy*, 2024.

[16] S. Rashed and C. Ozcan, "A new two-layer deep learning phishing and spam email detection model," *Electronics*, 2025.

[17] R. M. Gahara, A. Hidri, O. Arfaoui, and M. S. Hidri, "Large language models and convolutional networks as hierarchical multimodal robust spam detectors," *Procedia Computer Science*, 2025.

[18] S. A. Khan, K. Iqbal, N. Mohammad, R. Akbar, S. S. A. Ali, and A. A. Siddiqui, "A novel fuzzy-logic-based multi-criteria metric for performance evaluation of spam email detection algorithms," *Applied Sciences*, vol. 12, no. 14, p. 7043, 2022.

[19] M. Adnan, M. O. Imam, M. F. Javed, and I. Murtza, "Improving spam email classification accuracy using ensemble techniques: A stacking approach," *International Journal of Information Security*, vol. 23, no. 1, pp. 505–517, 2024.

[20] K. Iqbal, M. Khalid, S. Akhtar, S. Yasin, N. Ahmed, and A. Shahid, "Improving spam detection for German users: A machine learning approach to German email classification," *Kashf Journal of Multidisciplinary Research*, vol. 2, no. 6, pp. 81–99, 2025.

[21] P. L. R. Kameswari and P. S. Rani, "Multilingual spam classification using advanced deep learning techniques," in *Proc. Int. Conf. Sustainable Communication Networks and Applications (ICSCNA)*, Dec. 2024, pp. 1595–1601.

[22] P. Pantel and D. Lin, "SpamCop: A spam classification and organization program," in *Proc. AAAI-98 Workshop on Learning for Text Categorization*, Madison, WI, USA, Jul. 1998, pp. 95–98.

[23] E. H. Tusher, M. A. Ismail, and A. F. Mat Raffei, "Email spam classification based on deep learning methods: A review," *Iraqi Journal for Computer Science and Mathematics*, vol. 6, no. 1, p. 2, 2025.

[24] B. Yu and Z. B. Xu, "A comparative study for content-based dynamic spam classification using four machine learning algorithms," *Knowledge-Based Systems*, vol. 21, no. 4, pp. 355–362, 2008.

[25] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Next-generation spam filtering: Comparative fine-tuning of LLMs, NLPs, and CNN models for email spam classification," *Electronics*, vol. 13, no. 11, p. 2034, 2024.

[26] M. H. Alsuwit, M. A. Haq, and M. A. Aleisa, "Advancing email spam classification using machine learning and deep learning techniques," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 14994–15001, 2024.

[27] F. Jánez-Martino, E. Fidalgo, S. González-Martínez, and J. Velasco-Mata, "Classification of spam emails through hierarchical clustering and supervised learning," *arXiv preprint* arXiv:2005.08773, 2020.

[28] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email classification research trends: Review and open issues," *IEEE Access*, vol. 5, pp. 9044–9064, 2017.

[29] S. Rojas-Galeano, "Zero-shot spam email classification using pre-trained large language models," in *Proc. Workshop on Engineering Applications*, Cham, Switzerland: Springer Nature, Oct. 2024, pp. 3–18.

[30] L. Wang, "Spam email detection using Naïve Bayes classifier," in *ITM Web of Conferences*, vol. 70, p. 04028, 2025.

[31] H. Iswanto, E. Seniwati, Y. Astuti, and D. Maulina, "Comparison of algorithms on machine learning for spam email classification," *IJISTECH (International Journal of Information System and Technology)*, vol. 5, no. 4, pp. 446–455, 2021.

[32] P. Krishnamoorthy, M. Sathiyanarayanan, and H. P. Proença, "A novel and secured email classification and emotion detection using hybrid deep neural network," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 44–57, 2024.

[33] P. Bhargavi, "Spam email detection using machine learning and deep learning techniques," *International Journal of Research Publication and Reviews*, vol. 3, no. 11, pp. 1349–1352, 2022.

[34] Z. Zhang, Z. Deng, W. Zhang, and L. Bu, "MMTD: A multilingual and multimodal spam detection model combining text and document images," *Applied Sciences*, vol. 13, no. 21, p. 11783, 2023.

[35] S. A. Shah, E. A. Arputham, A. Ahmed, M. B. Farah, A. Shah, and A. Aziz, "Sorting the digital stream: Big data-driven insights into email classification for spam and ham detection," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Dec. 2023, pp. 5598–5607.