

Article

Adaptive Semi-Personalized Email Classification Model (ASPEC) with Incremental Learning

Worawit Kitikusoun and Nawaporn Wisitpongphan *

Department of Digital Network and Information Security Management, Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand; s6507031910026@email.kmutnb.ac.th

* Correspondence: nawaporn.w@itd.kmutnb.ac.th

Abstract

The volume of daily email traffic continues to grow rapidly, creating challenges in efficiently distinguishing important from irrelevant messages. Beyond spam detection, modern email systems classify messages into categories such as promotions, social, updates, and forums, many of which are ignored or deleted without review. To address this issue, researchers have explored intelligent classification systems to predict the importance of emails, enhance user productivity, and improve organizational communication efficiency. This study proposes an email classification model that adapts to different users' work functions and communication patterns within an organizational context. Using three-month historical real corporate anonymized email data from 9788 individuals across 12 work functions, the proposed Adaptive Semi-Personalized Email Classification Model (ASPEC) automatically retrieves each employee's occupational profile—including job category and years of work experience—from the organization's Human Resources (HR) system, enabling seamless personalization without manual configuration. ASPEC significantly improves email classification accuracy over the best-performing baseline of 73.50%, with incremental learning further enabling continuous adaptation to evolving data streams and achieving accuracy up to 92.57% in stable user segments. Unlike most existing email classification frameworks, which rely on static batch-learning models and lack memory-based or incremental update mechanisms, ASPEC addresses this gap by continuously adapting to evolving communication patterns without requiring full model retraining. The adoption of this incremental learning framework offers tangible benefits for organizations, including reduced manual email filtering workload, improved communication efficiency, and decreased operational burden on IT departments in managing email-related tasks and issues.

Keywords: email classification; feature selection; incremental learning; work-related email



Academic Editor: Zhiwen Yu

Received: 6 March 2026

Revised: 25 May 2026

Accepted: 26 May 2026

Published: 29 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The global volume of email usage has continuously increased in recent years. According to Statista [1], more than 347 billion emails are sent worldwide every day, and this number continues to grow. However, a large proportion of these emails are work-related and have varying levels of importance. When such emails are not properly managed or prioritized, users may spend a considerable amount of time filtering large volumes of messages, which negatively affects work productivity, time management, and the overall efficiency of enterprise communication.

Modern email systems, such as Gmail, categorize incoming messages into Primary, Promotions, Social, Updates, and Forums [2] to help users prioritize important emails and reduce the time spent on manual filtering, thereby improving productivity and communication efficiency.

This study proposes a semi-personalized email classification model that leverages occupational and email-usage behavioral contexts as key factors for improving classification accuracy. This model incorporates user-specific attributes, including job category, work experience, and email volume, to effectively distinguish work-related emails from non-work-related messages that do not require user attention. This approach enables users to prioritize essential emails more efficiently, thereby improving productivity and supporting effective organizational email management. A key practical feature of ASPEC is its integration with the organization's HR system, from which each employee's job category and work experience are automatically retrieved at classification time. This eliminates the need for manual user profiling and enables the system to be deployed with minimal configuration overhead in enterprise environments. Despite the growing body of research on email classification, most existing studies have focused primarily on spam detection, message-source identification, or semantic intent classification. Relatively little attention has been paid to the explicit distinction between work-related and non-work-related emails within organizational contexts—a classification task that is directly tied to employee productivity and enterprise communication efficiency. In particular, prior work has not sufficiently explored how occupational roles, work experience, and email usage behavior can be jointly leveraged to support this distinction. This gap motivates the present study and forms the central hypothesis: that a semi-personalized, group-level classification approach incorporating occupational and behavioral context will significantly outperform generic baseline models in work-related email identification.

To address the limitations of traditional classification models trained on generic datasets, which often fail to capture user-specific communication patterns, this study introduces an Adaptive Semi-Personalized Email Classification Model (ASPEC) that integrates machine learning techniques with adaptive feature selection and an incremental learning mechanism to construct optimized models for different user groups. When new emails arrive, the system selects the most effective model based on the recipient's occupational and behavioral contexts. The models for each group were trained and continuously refined using anonymized real-world corporate email data. This approach enhances classification accuracy, reduces feature redundancy, and improves the adaptability of the email management system in dynamic organizational environments. Importantly, ASPEC is designed to be lightweight and cost-effective, requiring only classical machine learning algorithms and email metadata—without reliance on large language models (LLMs), deep learning infrastructure, or AI agent frameworks. This makes it particularly well-suited for deployment in small- to medium-sized enterprises (SMEs) that require an internally manageable and privacy-preserving email classification solution.

The key contributions of this study are fourfold. First, we introduce an underexplored email classification task—distinguishing work-related from non-work-related emails in an organizational context—that has received limited attention in the existing literature. Second, we propose a three-dimensional adaptive group-level architecture that implicitly functions as an intra-ensemble model selection strategy tailored to occupational and behavioral user profiles. Third, the framework is designed to operate entirely on email metadata without accessing body content, thereby supporting privacy-preserving deployment in organizational environments. Fourth, the proposed system is validated on a large-scale real corporate dataset comprising 423,043 emails from 9788 employees across 12 functional departments, providing empirical evidence of its practical applicability.

2. Background

Email classification relies on the extraction of informative features from email data, which are transformed into numerical representations for machine learning processes. Commonly used features can be categorized into several groups based on the nature of the data and the intended application. Each feature group plays a critical role in determining the accuracy and effectiveness of the email classification models.

In this study, the classification models considered are classical machine learning algorithms, specifically Logistic Regression (a weak linear learner), Decision Tree, and Random Forest (strong ensemble learner). Deep learning models and large language models (LLMs) were intentionally excluded, as the primary objective of this work is to construct an interpretable, computationally efficient, and privacy-preserving system suitable for deployment in small- to medium-sized organizational environments.

Email Structural Features are derived from the email header and do not expose the internal content of the message. As a result, they are particularly suitable for organizational environments where data privacy is a major concern. Features in this category include the sender address, sender domain, sender status (internal or external), email type (e.g., new, reply, or forward), email size, and the presence or absence of attachments. These structural features effectively capture communication patterns and email usage characteristics without relying on the email body [3–5].

Temporal Features (also referred to as time-based or dynamic features) are used to describe user communication behavior over time and are closely related to organizational working contexts. Unlike static structural or user-context attributes, these features are inherently time-varying and capture behavioral patterns that shift across hours, days, and organizational cycles—making them particularly challenging for traditional batch-trained ML models that assume a fixed feature distribution. Examples of temporal features include the day of the week on which emails are sent or received, the distinction between working and non-working hours, and the frequency of email reception within a given time period. These time-based signals are particularly valuable because work-related emails are predominantly concentrated during business hours and weekdays, whereas non-work-related messages such as promotions and newsletters tend to arrive outside these windows. Temporal features, therefore, enable the classifier to exploit behavioral regularities that are invisible to classic machine learning models trained solely on content or structural attributes, resulting in improved separation of work-related from non-work-related emails without relying on the email body content [3,6].

Behavioral Features focus on reflecting the email usage patterns of individual users or user groups and are essential for developing semi-personalized or personalized classification models. These features include the volume of emails received over a specific period, volume grouping, reply and forwarding patterns, and historical email usage during working hours. Incorporating behavioral features enables models to learn user-specific characteristics more effectively than those trained solely on generic data [7].

User Context Features represent user-specific characteristics related to occupational roles and work experience that influence the types and patterns of emails received. Features in this group include job category, length of work experience, and position. By integrating user context features, email classification can be better aligned with users' responsibilities and organizational roles, allowing for more accurate prioritization of important emails [7].

Lightweight Content Features are the only textual features in the proposed framework and are employed to enrich the semantic context of emails while preserving data privacy. Unlike the structural, temporal, behavioral, and user context features—which are all quantitative in nature—lightweight content features are derived from the email subject line using Term Frequency–Inverse Document Frequency (TF-IDF) vectorization. To ensure

language independence, email subjects written in Thai are first translated into English prior to TF-IDF processing, allowing the semantic representation to remain consistent regardless of the original language of the email. This approach enriches classification performance without compromising the confidentiality of organizational information [8,9].

3. Related Work

Research on email classification has evolved continuously over the past two decades to address the rapid growth in email volume, as well as the increasing diversity and complexity of message content. Early approaches relied primarily on rule-based systems defined by expert-crafted heuristics. As labeled datasets and computational resources became more widely available, machine learning techniques were introduced to improve classification accuracy and generalization capability. Subsequently, deep learning methods have gained prominence due to their ability to learn semantic structures and contextual representations from email data more effectively. In more recent studies, incremental and adaptive learning approaches have attracted increasing attention because they are better suited to handling temporal changes and concept drift in real-world email streams. Accordingly, this section reviews prior work on email classification across different methodological phases and discusses their limitations, which motivate the need for more flexible and context-aware classification approaches.

3.1. Rule-Based/Signature Filtering

In the early stages of unsolicited email filtering, most systems relied on rule-based and signature-driven methods grounded in the email structure defined by Request for Comments (RFC) 5322 [3]. In this context, email metadata refers to the structured, non-content descriptive attributes embedded in the email header, including fields such as the sender address, recipient address, timestamp, subject line, message size, routing information (e.g., IP addresses and mail server hops), and protocol-level identifiers (e.g., Message-ID). These metadata fields provide rich signals for classification without exposing the private body content of the message and form the foundational basis for the feature engineering approach adopted in this study. These approaches used manually crafted conditions such as blacklist/whitelist rules, suspicious keywords (e.g., “verify your password,” “account suspended”), and signature patterns to identify spam-like behavior. Cormack [10] summarized the progression of early filters from hand-crafted rules to the adoption of linear classifiers—such as Naïve Bayes and Logistic Regression—which model the decision boundary as a linear function of the input features to improve accuracy. These linear classifiers represented the first generation of data-driven email filtering, offering improved generalization over purely rule-based systems while remaining computationally tractable. Wong and Schlitt [4] defined the Sender Policy Framework (SPF) standard (RFC 4408), enabling domain owners to specify authorized mail servers and thereby building sender-authentication systems that effectively reduce spoofed emails. Allman et al. further strengthened email authentication by introducing DomainKeys Identified Mail (DKIM), which uses public-key cryptography to digitally sign messages and verify sender identity [5].

A milestone in this era is the work by Sahami et al. [11], who integrated rule-based features such as suspicious keywords, heuristic patterns, and email-header metadata with a Naïve Bayes (NB) classifier, achieving significantly improved accuracy. Similarly, Nguyen et al. [12] proposed the Dynamic Sender Policy Framework (DSPF) approach to validate sender domains, introducing a dynamic rule capable of detecting domain spoofing and reducing false positives.

While rule-based and signature-driven methods established the foundational principles of email filtering, their reliance on manually crafted rules limited their scalability and adaptability to evolving spam patterns. This limitation motivated a paradigm

shift toward data-driven, machine learning-based strategies, which are discussed in the following section.

3.2. Machine Learning Era

The emergence of Machine Learning (ML) has led to a major leap in spam filtering capabilities. Popular algorithms include Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Neural Network (NN), and various hybrid techniques.

Ajani and Ferrante reported that RF achieved 95.87% accuracy, whereas the hybrid RF–NN–NB model achieved 97.22% [13]. Awad and Elseuofi found that NB achieved 99.46% accuracy, while their proposed Rough Sets classifier achieved 99.72%, outperforming the Artificial Immune System (AIS) and SVM [14].

A landmark contribution is the work by Androutsopoulos et al. [15], who evaluated NB models under multiple configurations and introduced the Total Cost Ratio (TCR), a cost-sensitive evaluation metric designed to reflect the high cost of false positive errors in spam filtering using the Ling-Spam dataset, a publicly available benchmark corpus. Aslantaş et al. applied the Binary Firefly Algorithm (BFA) for feature selection, reducing the number of features from 57 to 21 and improving the accuracy of the NB classifier from 79.6% to 95.14% [8]. Christina et al. compared Multilayer Perceptron (MLP), J48 decision tree (an implementation of the C4.5 algorithm), and NB, finding that MLP achieved the highest accuracy of 99.3% [16].

3.3. Deep Learning Era

Driven by large datasets and enhanced computational power, Deep Learning (DL) has become prominent. Models such as convolutional neural networks (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and attention-based architectures are capable of effectively capturing semantic and contextual features. It is worth noting that CNN architectures are primarily suited for extracting local, stationary patterns from fixed-dimensional feature spaces—analogue to their application in image recognition—whereas recurrent architectures such as LSTM and GRU are specifically designed to model sequential dependencies and temporal dynamics in text and time-series data. In the context of email classification, RNN-based models are therefore more naturally aligned with capturing the sequential structure of email content and temporal communication patterns than CNN-based models.

AbdulNabi and Yaseen reported that bidirectional encoder representations from transformers (BERT) [17] base case achieved 98.67% accuracy, outperforming Bidirectional Long Short-Term Memory (BiLSTM) and classical ML models [18]. Muhammad et al. developed a CNN–LSTM hybrid model, reaching an accuracy of 98.68% [19]. Nasreen et al. proposed a gray wolf optimizer (GWO)–optimized BERT model incorporating CNN, LSTM, and BiLSTM components, achieving 99.14% accuracy on the Ling-Spam dataset [9]. Pan et al. introduced a Semantic Graph Neural Network (SGNN), which improved the classification accuracy over CNN while reducing memory usage [20]. Halder et al. demonstrated that CNN and BERT achieved the best performance among ML, DL, and Large Language Models (LLMs), which are transformer-based models pretrained on large-scale corpora [21]. Tida and Hsu proposed the Universal Spam Detection Model (USDm), which achieved an accuracy of 97% and an F1-score of 0.96 [22].

Despite these impressive accuracy gains, a fundamental limitation of deep learning models is their decreasing interpretability as architectural complexity increases. Models such as BERT, CNN–LSTM hybrids, and GNN-based classifiers operate as black-box systems, making it difficult for practitioners to understand or audit the classification

decisions—a critical concern in organizational settings where transparency, accountability, and regulatory compliance are required. This explainability-accuracy trade-off is a key motivation for the present study's adoption of interpretable classical machine learning classifiers within the ASPEC framework.

3.4. Incremental/Adaptive Learning Era

As email patterns change rapidly, batch-learning models degrade over time. Incremental and adaptive learning models update continuously and effectively to handle concept drift. In this study, incremental learning was implemented in a batch-incremental manner, where models were periodically updated using newly accumulated email data rather than through instance-by-instance updates. Elwell and Polikar introduced Learn++.NSE (Learn++ for Non-Stationary Environments), an incremental ensemble that retains useful knowledge while forgetting outdated behaviors [23]. Sheu et al. proposed a spam filtering approach based on email header information combined with an incremental learning mechanism. Performance evaluation metrics were employed to assess the system, demonstrating that incremental learning can progressively improve the accuracy and correctness of email classification as new data are incorporated into the learning process [6]. Mang et al. proposed an incremental learning based framework integrated with ML ensembles for image spam filtering, demonstrating that such approaches can achieve high classification performance with a low false positive rate [24].

3.5. New Email Classification Paradigm and Contribution

Recent studies on email classification have extended beyond traditional spam and ham detection to address fine-grained and context-aware categorization tasks. For example, Kang et al. [25] proposed a deep learning framework for automatically distinguishing human-generated emails from machine-generated messages. Their approach integrates multiple CNN-based sub-models to analyze email content, sender attributes, and user interaction patterns, achieving superior performance compared to baseline classifiers and state-of-the-art language models. Notably, the proposed framework has been deployed in large-scale production environments, demonstrating its practical applicability in real-world email systems where automated and system-generated messages constitute a substantial portion of the overall traffic.

In a different line of research, Shu et al. [26] investigated email intent detection using weak supervision, aiming to classify emails based on their underlying communicative intent rather than on surface-level labels. To address the high cost and privacy concerns associated with large-scale manual annotation, their method leverages user interaction logs as a source of weak supervision together with a limited set of clean labeled data. The authors propose an end-to-end deep neural network with a self-paced learning mechanism to effectively integrate noisy weak signals with reliable annotations. The experimental results show that this approach achieves robust performance across multiple email-intent classification tasks. It should be noted that the self-paced learning mechanism employed by Shu et al. differs fundamentally from the incremental and adaptive learning approach adopted in the present study. Self-paced learning is a curriculum-based training strategy that controls the order and weighting of training samples within a static dataset, progressing from easier to more difficult instances. It is not designed to handle continuously arriving data streams or concept drift over time. By contrast, the incremental learning mechanism in ASPEC periodically retrains group-level models using newly accumulated email data, enabling the system to adapt to evolving communication patterns in real-world organizational environments.

Despite the growing body of research on email classification, ranging from spam filtering and human versus machine generated email detection to intent-based categorization, most existing studies have focused primarily on message source characteristics, semantic intent, or user interaction behaviors. However, relatively little attention has been paid to the explicit classification of work versus non-work-related emails within organizational contexts. In particular, prior studies have not sufficiently explored how occupational roles, work experience, and email usage behavior can be jointly leveraged to distinguish emails that are essential for work from those that do not require immediate attention. This research gap motivates the present study, which aims to develop a semi-personalized email classification approach that explicitly targets work-related email identification to support effective prioritization and enhance productivity in enterprise email management.

It is acknowledged that defining email importance is inherently subjective and dynamic. Different users—even within the same job role—may assign varying levels of importance to the same message, and an individual's priorities can shift over time depending on ongoing projects, deadlines, and organizational context. To mitigate this subjectivity, this study deliberately reframes the classification objective from predicting personal importance to distinguishing work-related from non-work-related emails—a more operationally consistent and group-level verifiable label. Nonetheless, the dynamic, user-level nature of email relevance remains an open challenge and is identified as a direction for future personalization research.

4. Methodology

This section presents the methodological process for constructing and evaluating the Adaptive Semi-Personalized Email Classification Model (ASPEC) with Incremental Learning. The overall workflow includes three key stages: dataset collection, data preprocessing, and data cleaning and transformation for machine learning applications.

4.1. Email Dataset

In this phase, a dataset is constructed from the email server logs of a large private enterprise in the financial services sector in Thailand, employing more than 9000 staff across 12 functional departments. The organization's identity remains confidential in accordance with the data-sharing agreement. The collected logs consisted of email transactions within the organization's internal system. Before data use, all personally identifiable information (PII) was removed or anonymized to comply with the Personal Data Protection Act (PDPA) and to ensure the confidentiality of user communications. The use of this dataset was authorized by the organization's management in accordance with Thailand's Personal Data Protection Act (PDPA B.E. 2562). All data were fully anonymized prior to access by the research team, and no individual user could be identified from the retained metadata. As only non-identifiable email header attributes were utilized, no individual informed consent from employees was required under the applicable regulatory framework. These procedures ensured that the research was conducted with full respect for user privacy and organizational data governance policies.

To prevent exposure to personal or business-sensitive content, only non-identifiable metadata was retained. The data were organized by month and covered a three-month period consisting of 625,069 emails in total. The first two months comprised 423,043 emails, which were used for model training and testing. The third month comprised 202,026 emails, which were reserved exclusively for incremental learning evaluation. This dataset provides a reliable basis for analyzing organizational communication patterns and employee email usage behaviors with high accuracy, without compromising privacy.

4.2. Data Preprocessing

In this step, the same anonymized dataset collected from the private company's email usage logs was used. The dataset includes only email header information, such as the Subject, Sender, Date/Time, and other technical attributes (for example, file size, source, and destination domains). These fields cannot directly identify individuals but retain analytical value for studying communication patterns and organizational email behavior, serving as a foundation for model development.

Following the Internet Message Format standard (RFC 5322) [3], irrelevant elements such as HTML tags, URLs, emojis, and special symbols have been removed to improve data consistency. All text data were normalized and encoded in UTF-8 to support both English and Thai content.

It is worth noting that Thai is a Subject–Verb–Object (SVO) language, sharing a similar syntactic structure with English. This structural similarity facilitates translation and semantic alignment between the two languages. However, Thai text presents unique preprocessing challenges, including the absence of explicit word boundaries and the lack of inflectional morphology. To address these challenges and ensure consistent feature extraction, all Thai email subject lines were translated into English prior to TF-IDF vectorization, thereby standardizing the linguistic representation across the dataset. The dataset contains email subjects written in both Thai and English, reflecting the bilingual communication environment of the organization. A fine-grained comparison of the relative informativeness of Thai versus English subject features was not conducted in the present work but is identified as a promising direction for future research, particularly for organizations operating in multilingual contexts.

4.3. Email Data Preparation and Cleaning

Data preparation and cleaning were performed to ensure that the dataset was accurate, complete, and suitable for machine learning algorithms. The process comprises the following stages:

1. Duplicate Removal: Emails were sorted by timestamp and verified using the unique Message-ID field to eliminate duplicate records and prevent redundant learning.
2. Text Cleaning: Non-informative characters such as special symbols, HTML tags, URLs, and emojis were removed using regular expression-based pattern matching.
3. Irrelevant component removal: System-generated header or footer sections were excluded to retain meaningful textual content.
4. Class labeling: Each email was manually labeled by two independent members of the research team as either Work-related or Non-work-related, following a predefined annotation guideline. To assess labeling consistency, a random sample of 200 emails was independently annotated by both annotators without prior discussion. Emails were classified as Work-related if the subject line contained explicit references to organizational tasks, meetings, reports, projects, system notifications, or internal communications. Emails were classified as Non-work-related if the subject indicated promotional content, newsletters, personal messages, or non-task system-generated notifications. In cases of disagreement between annotators, a third senior researcher served as the final arbitrator. Inter-annotator agreement was assessed using Cohen's Kappa coefficient on the 200-sample subset, yielding a score of 0.80, which indicates strong labeling consistency [27]. The remaining emails were subsequently labeled by a single annotator following the same established guidelines.

It is acknowledged that subject line-based labeling may not capture all work-related signals present in the email body and that certain emails with ambiguous or generic subject lines may be more challenging to classify consistently. However, this approach

was necessitated by organizational data privacy constraints (PDPA), which prohibited access to email body content. Subject lines have been shown in prior work to carry sufficient discriminative value for coarse-grained email categorization [6], and the strong inter-annotator agreement observed in this study ($\kappa = 0.80$) supports the reliability of the resulting labels within this constraint.

5. Data transformation: The cleaned text was processed using:

- Text normalization: Converting all text to lowercase to reduce redundancy.
- Tokenization and stopword removal: Words were segmented, and common stopwords (e.g., the, is, and) were removed using standard natural language processing procedures [28].
- Stemming/lemmatization: Reducing words to their root form (e.g., running to run) to decrease dimensionality [28].
- Vectorization: Transforming text into numerical feature vectors using Term Frequency-Inverse Document Frequency (TF-IDF) or count vectorization methods [29].

The TF-IDF weight of term t in document d within corpus D is computed as

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log\left(\frac{|D|}{1 + |\{d \in D : t \in d\}|}\right) \quad (1)$$

where $\text{TF}(t, d)$ denotes the number of occurrences of term t in the subject line of email d , $|D|$ is the total number of training emails used to fit the TF-IDF vectorizer, and $|\{d \in D : t \in d\}|$ is the number of training emails containing term t . Higher TF-IDF weights are assigned to terms that are frequent in a specific email but rare across the training corpus, thereby capturing semantically discriminative signals in the email subject line.

While these transformation steps involve a degree of controlled information reduction—for instance, stemming may collapse semantically distinct word forms, and stopword removal eliminates function words—this trade-off is intentional. The primary objective is to reduce feature dimensionality and suppress noise, thereby improving model generalization. Given that only email subject lines (rather than full body text) are used as the textual input, the resulting information loss is minimal and does not significantly compromise the semantic discriminability of the features.

In the email data preparation and cleaning stage, this study selects a set of features derived from email data that do not reveal internal message content to comply with data privacy constraints within organizational environments. The selected features include email size and sender status, indicating whether the sender is internal or external to the organization. In addition, temporal features are incorporated, such as the day of the week and whether emails are received during working or non-working hours, to reflect communication patterns associated with actual work activities. The feature set also covers behavioral and contextual user information, including the length of employment, work-experience groupings, job category or role, and the volume of emails received within a given period. Furthermore, limited semantic information derived from translated email subject lines is utilized to enhance contextual understanding without relying on the full email content. Prior to model construction, all data underwent cleaning and transformation processes, such as converting categorical attributes into numerical representations and normalizing data formats to ensure suitability for feature selection and machine-learning procedures. These steps support the development of the Adaptive Semi-Personalized Email Classification (ASPEC) model, enabling accurate classification performance that aligns with real-world organizational email usage.

Regarding class distribution, the labeled dataset contains approximately 56.5% work-related and 43.5% non-work-related emails represent a mild class imbalance. Given this moderate ratio, no resampling technique (e.g., Synthetic Minority Oversampling Technique (SMOTE) or undersampling) was applied. Instead, F1-score, precision, and recall were reported alongside accuracy to ensure a balanced evaluation of model performance across both classes. To prevent data leakage, TF-IDF vectorization was fitted exclusively on the training partition and subsequently applied to the test partition without refitting. A stratified 70:30 train-test split was applied within each user group to preserve the original class distribution across both partitions.

4.4. Baseline Performance

The Decision Tree classifier employed in this study is based on the Classification and Regression Tree (CART) algorithm, implemented using the Gini impurity criterion. This is distinct from ensemble tree-based methods such as Random Forest or XGBoost, which aggregate multiple trees. The CART-based single Decision Tree was selected as a baseline to represent interpretable, non-ensemble tree classifiers.

The Gini impurity at a node is defined as:

$$I_G(p) = 1 - \sum_{i=1}^J p_i^2 \quad (2)$$

where J is the number of classes and p_i is the proportion of samples belonging to class i at the node. A Gini impurity of zero indicates a pure node containing samples of a single class, whereas higher values reflect greater class mixing. The CART algorithm selects the feature and threshold that minimize the weighted Gini impurity across the resulting child nodes at each split.

The baseline performance evaluation was conducted using labeled data to establish a reference benchmark for comparing the performance of different classification models. The selection of baseline models was guided by prior comparative studies that systematically evaluated multiple machine learning algorithms under standardized experimental conditions for email spam classification.

Omotehinwa and Oyewola [30] conducted a comparative evaluation of ensemble machine learning algorithms—including Random Forest (RF) and XGBoost—with hyperparameter optimization via grid-search cross-validation on the Enron dataset. Their experimental results demonstrated that RF and XGBoost consistently achieved high classification accuracy and robustness, whereas Logistic Regression exhibited competitive performance and strong suitability for high-dimensional textual data. Similarly, Yusupov et al. [31] performed a comprehensive comparison between classical machine learning models and deep learning approaches for email spam classification using both TF-IDF and word embedding representations. Their findings indicated that RF achieved the best performance among traditional machine learning classifiers, particularly when TF-IDF features were employed, whereas LR and DT maintained stable and competitive results across different feature representations, making them appropriate baseline models. Furthermore, Kontsewaya et al. [32] evaluated the effectiveness of several machine learning methods for spam detection under identical experimental conditions. Their results showed that RF achieved the highest overall accuracy, whereas LR and DT provided an effective balance between classification performance and model complexity, supporting their applicability in practical email filtering systems.

Based on these comparative studies, LR, DT, and RF were selected as baseline classifiers because they consistently demonstrated effective and reliable performance in email spam classification tasks. While organizational email data may exhibit complex,

multi-dimensional interaction patterns that challenge simple classifiers, ASPEC addresses this complexity through its adaptive feature selection and group-level segmentation strategy rather than through deep learning architectures. By identifying the optimal feature subset for each user group independently, the effective dimensionality and complexity of each sub-problem are substantially reduced, enabling classical ML classifiers to achieve competitive performance while maintaining interpretability and computational efficiency suitable for organizational deployment.

The baseline performance followed the workflow illustrated in Figure 1, in which each model was trained and evaluated using the same dataset split, with 70% of the data allocated for training and 30% allocated for testing performance.

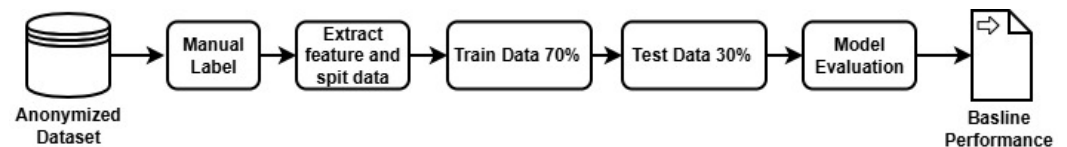


Figure 1. Overall workflow of evaluating baseline email classification performance.

Standard performance metrics, including accuracy, precision, recall, and F1-score, were computed to ensure a consistent comparison among all evaluated models.

A supervised feature selection approach was employed using labeled data to evaluate the baseline performance of the classification models under uniform experimental conditions. The objective of this evaluation was to identify the model that achieved the highest accuracy and overall performance among all the tested algorithms. The baseline performance results obtained from the labeled dataset are summarized in Table 1.

Table 1 presents the baseline email classification results derived from the labeled dataset. The experimental results indicate that the Random Forest algorithm achieved the best overall performance when considering accuracy and balanced performance metrics, with an accuracy of 72.58%, an F1-score of 73.58%, a precision of 70.96%, and a recall of 76.37%. These results demonstrate the ability of Random Forest to maintain an appropriate balance between precision and recall, resulting in a relatively low classification error rate.

For the Decision Tree algorithm, the results showed an accuracy of 73.50% and an F1-score of 73.37%, with precision and recall values of 71.37% and 75.47%, respectively. Although these values are not the highest among all the models, they remain competitive, indicating that a single-tree model can effectively learn fundamental data patterns, even without prior feature selection.

In contrast, Logistic Regression produced the lowest performance among the evaluated models, achieving an accuracy of 71.19%, an F1-score of 66.52%, a precision of 63.85%, and a recall of 69.40%. These results suggest that linear models may be insufficient for capturing the complex interactions and non-linear patterns present in email classification tasks.

Table 1. Baseline test results.

Algorithm	Accuracy	F1-Score	Precision	Recall
Random Forest	72.58%	73.58%	70.96%	76.37%
Decision Tree	73.50%	73.37%	71.37%	75.47%
Logistic Regression	71.19%	66.52%	63.85%	69.40%

5. Adaptive Semi-Personalized Email Classification (ASPEC)

The proposed ASPEC is a multi-level, feature-based classification architecture that integrates user-specific behavioral segmentation and adaptive feature selection to

improve classification accuracy across different organizational contexts. Specifically, the architecture operates across three independent segmentation dimensions—job category, work experience, and email volume—each of which is associated with its own pre-determined optimal feature subset and best-performing classifier, identified through systematic filter and wrapper feature selection processes. For each incoming email, the system evaluates all three dimension-level models and selects the one yielding the highest accuracy for the recipient's profile, as mathematically defined in Equation (4). Figure 2 illustrates the overall concept of ASPEC.

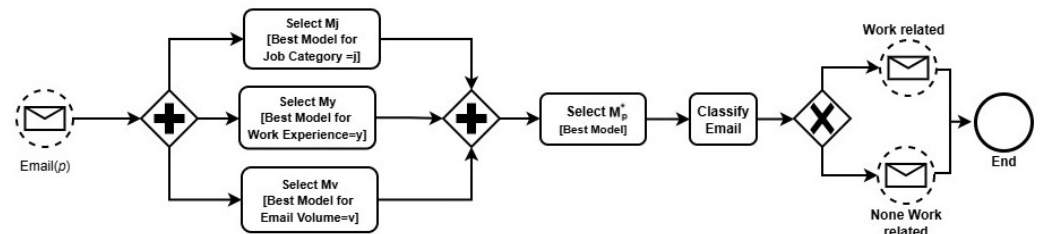


Figure 2. Adaptive Semi-Personalized Email Classification Model.

ASPEC can be interpreted as an implicit stratified ensemble architecture. Each of the three segmentation dimensions—job category, work experience, and email volume—independently contributes a trained model, and the final classification model is selected based on the highest accuracy among the three candidates for each individual employee (Equations (4) and (5)). This model selection step functions analogously to an ensemble voting mechanism, albeit applied at the group-segmentation level rather than through bagging or boosting. The final classification layer, which assigns each email to either the work-related or non-work-related class, constitutes a parallelized binary decision step that operates uniformly across all selected models.

It is important to note that ASPEC adopts an intra-group learning strategy, in which each group model is trained independently using only the data associated with that specific group. No cross-group data sharing or inter-group knowledge transfer is performed. The binary classification labels (Work-related vs. Non-work-related) remain consistent across all groups.

5.1. User Grouping

To enable adaptive classification that reflects individual or group-level differences, the dataset was categorized into three primary user dimensions as follows:

- **Job Category:** Users were divided into 12 groups according to their job roles, representing variations in email characteristics between professional functions, as summarized in Table 2.
- **Work Experience:** Users were grouped into seven categories according to their years of experience, reflecting differences in email usage associated with tenure, as shown in Table 3.
- **Email Volume:** Users were classified into five groups based on the number of emails they received per week, representing different levels of communication frequency, as listed in Table 4.

The three segmentation dimensions are treated as mutually independent in ASPEC: each dimension defines its own group models without cross-dimensional dependencies. The grouping criteria were established as follows: job categories were defined in accordance with the organization's existing HR classification structure; work experience boundaries were set based on commonly recognized tenure milestones in human resource management practice; and email volume thresholds were determined empirically from the observed

distribution of weekly email counts in the dataset. This independence ensures that the model selection mechanism in Equation (5) evaluates three distinct, non-overlapping perspectives on each employee's communication profile.

Table 2. Job category groups.

Group	Job Description
J1	Senior Executives (President, EVP, SVP, VP, AVP)
J2	IT (Support, Network, Programmer, CyberSec)
J3	Sales & Marketing (Sales Mgr, Marketing, Direct Sales)
J4	HR (HRD, Talent, Recruit)
J5	Dev & Analysis (Data, Biz, Product)
J6	Finance & Accounting (Acct, Analyst, Billing)
J7	Customer Service (Call Center, Support)
J8	Technical & Engineering (Engineer, Support)
J9	Legal & Policy (Legal, Data Privacy)
J10	Others (Receptionist, Driver, Admin)
J11	Manager / Assistant Manager
J12	Supervisor

Table 3. Work experience groups.

Group	Years of Experience
Y1	Less than 4 years
Y2	4–10 years
Y3	10–15 years
Y4	15–20 years
Y5	20–25 years
Y6	25–29 years
Y7	More than 30 years

Table 4. Weekly email volume groups.

Group	Number of Emails (Per Week)
V1	≤ 500
V2	501–1000
V3	1001–2000
V4	2001–3000
V5	> 3000

5.2. Feature Selection

This study employed two feature selection techniques: the filter method and the wrapper method (forward selection) to identify informative feature subsets for each user group while preserving user privacy.

The feature set consisted of the email size category, email type, sender status, and reply/forward status, representing the structural characteristics of emails. Temporal communication patterns were captured through weekday and working-hour indicators. User context was represented by years of work experience, Work experience category, job category, and email volume group. A limited semantic context was incorporated through the translated email subject, in which the TF-IDF technique was applied to weight

important terms and capture the relative semantic importance of the words appearing in the email subject. This approach enhances the quantitative representation of semantic information, despite relying on textual content with a limited contextual scope. The role of TF-IDF in this context is to assign higher weights to subject-line terms that are distinctive and informative for classification—such as work-specific keywords (e.g., meeting, invoice, report, deadline)—while suppressing common terms that appear uniformly across both work-related and non-work-related emails. This weighting mechanism allows the classifier to focus on semantically discriminative signals within the subject line. It should be noted that a formal ablation comparison between TF-IDF and raw count vectorization was not conducted in the present study; this comparison is identified as a direction for future work to further quantify the contribution of the TF-IDF transformation to overall classification performance.

It is important to clarify that the textual feature extraction approach adopted in ASPEC is grounded in the text mining paradigm, which relies on statistical representations such as TF-IDF to quantify term importance across a document corpus. This is methodologically distinct from natural language processing (NLP) approaches—such as those employing BERT or recurrent neural networks—which model deep semantic structures and contextual word dependencies. The text mining approach was deliberately chosen in this study to maintain computational efficiency, interpretability, and compatibility with the classical ML classifiers employed in ASPEC.

A semi-personalized learning strategy was adopted in which models are customized at the user-group level, rather than for individual users. The groups are defined based on job role, email volume, and work experience, enabling a practical trade-off between classification performance and computational efficiency.

5.2.1. Filter Method

This method is computationally efficient and independent of any specific algorithm, making it suitable for the preliminary screening of large-scale datasets.

The Chi-Square statistical test [33] was applied to evaluate the strength of the association between each feature and the target variable, retaining features with statistically significant relationships to improve classification precision.

The Chi-Square statistic for evaluating the association between each feature and the target class is defined as:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

where O_i denotes the observed frequency, E_i denotes the expected frequency, and k is the number of cells in the contingency table. More specifically, we calculate the difference between observed and expected frequencies for every possible email scenario. Features with higher χ^2 values indicate stronger statistical dependence with the target class and are therefore retained for model training.

5.2.2. Wrapper Method (Forward Selection)

This method focuses on maximizing classification accuracy by iteratively testing the performance of the model with subsets of progressively expanded features [34]. Although computationally intensive, it provides superior accuracy because each subset is directly validated using the classification model. The Forward Selection technique evaluates one feature at a time, adding the feature that yields the greatest incremental improvement in classification accuracy until the optimal feature combination is achieved. Forward selection was preferred over backward selection because it is computationally more efficient when the number of candidate features is large relative to the group size, and it provides a naturally

interpretable ranking of feature contributions. The stopping criterion was defined as the point at which no additional feature further improved classification accuracy on the 30% holdout validation set. No probabilistic inclusion threshold or information criterion (e.g., Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)) was applied, as the primary optimization objective was the direct maximization of classification accuracy.

The integration of these two techniques enables the ASPEC model to effectively select appropriate features for each user group, allowing the model to adapt to the distinct characteristics and behavioral patterns of different user segments.

5.3. Model Selection

ASPEC selects the most appropriate classifier algorithm and feature set for each group categorized by job role, work experience, and email volume. This approach enabled the construction of an optimized model tailored to the characteristics of each group, as illustrated in Figure 2.

Let p denote an individual employee from the total of P employees. Let $\text{Email}(p)$ denote the set of all emails associated with employee p , where employee p belongs to the job category group j , work-experience group y , and weekly email-volume group v .

Let

- M_j denotes the best-performing model for job category group j .
- M_y denotes the best-performing model for work-experience group y .
- M_v denotes the best-performing model for the email-volume group v .

For the email of employee p , the Adaptive Semi-Personalized Email Classification (ASPEC) system selects the optimal model for that individual according to the following rule:

$$M_p^* = \begin{cases} M_j, & \text{if } A(M_j) = \max\{A(M_j), A(M_y), A(M_v)\}, \\ M_y, & \text{if } A(M_y) = \max\{A(M_j), A(M_y), A(M_v)\}, \\ M_v, & \text{if } A(M_v) = \max\{A(M_j), A(M_y), A(M_v)\}. \end{cases} \quad (4)$$

- $A(M_j)$ represents the highest classification accuracy of the model for the job category group j .
- $A(M_y)$ represents the highest classification accuracy of the model for the work-experience group y .
- $A(M_v)$ represents the highest classification accuracy of the model for the email-volume group v .

This selection rule is equivalent to selecting the model with the highest classification accuracy among the three candidate dimensions, which can be expressed compactly as:

$$M_p^* = \arg \max_{m \in \{M_j, M_y, M_v\}} A(m) \quad (5)$$

where $A(m)$ denotes the classification accuracy of model m , and M^* is the optimal model selected for employee p .

Upon receiving an incoming email, the system first identifies the intended recipient. The system then connects to the organization Human Resources (HR) system to determine the recipient's job category group and years of work experience. The weekly email volume of the recipient was maintained as an internal variable in the ASPEC system.

Subsequently, the performance of the candidate models was compared according to the model selection rule formalized in Equation (5) to determine the optimal model for each individual. The selected model was then applied to classify all emails associated with that employee as either work-related or non-work-related.

5.4. Incremental Learning

The Incremental Learning process, as illustrated in Figure 3, demonstrates how the system can continuously update and refine its model to adapt to newly available data.

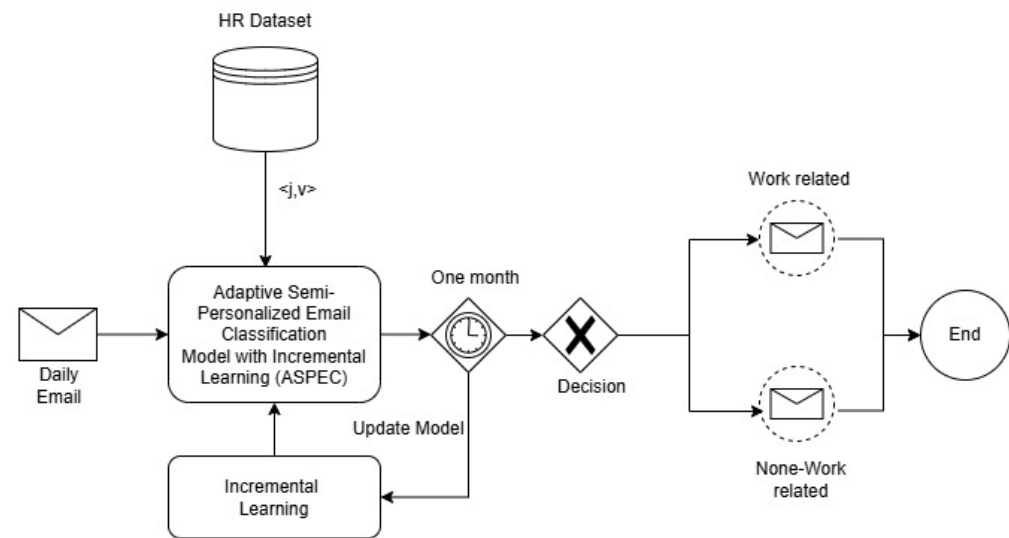


Figure 3. Incremental data learning process for ASPEC.

The nature and types of emails received may vary over time, depending on organizational activities and contextual factors. ASPEC must constantly update its model through an Incremental Learning process to ensure that it remains current and accurate. This process involves retraining the models using newly classified emails from ASPEC and incorporating feedback on misclassified emails received from each user since the last model update. These data were used as inputs for training and testing to identify the most suitable updated model.

If the updated model demonstrates improved classification performance compared with the previous version, ASPEC replaces the existing model with the updated version; otherwise, the previous model is retained, as formalized in Equation (7).

Once the optimal model is obtained, the system applies it to classify the incoming emails. This approach enables the model to dynamically adapt to changes in real-world email usage behaviors while maintaining consistent accuracy and effective performance.

Let the model be updated at time T_k , where $k = 1, 2, \dots$. The update interval was defined as $u = T_k - T_{k-1}$. The set of new emails belonging to group g received during the interval (T_{k-1}, T_k) is expressed as follows:

$$E^{(g,k)} = \{e_1^{(g,k)}, e_2^{(g,k)}, \dots, e_{N^{(g,k)}}^{(g,k)}\} \quad (6)$$

where $N^{(g,k)}$ represents the total number of new emails in group g during the k -th model update cycle. For each group, the emails were randomly split into 70% for training and 30% for testing to obtain the best-performing model $M^{(g,k)}$.

The new model for each group was determined by comparing the accuracy of the updated model with that of the previous one, as defined by the following equation:

$$M^{(g,k)} = \begin{cases} M^{(g,k)}, & \text{if } A(M^{(g,k)}) \geq A(M^{(g,k-1)}), \\ M^{(g,k-1)}, & \text{if } A(M^{(g,k)}) < A(M^{(g,k-1)}). \end{cases} \quad (7)$$

This Incremental Learning process enables the ASPEC to learn from new data in real time without the need for full retraining. This ensures that the model maintains stable accuracy and adaptability in dynamic, constantly changing environments.

6. Results

6.1. Best Models for Each Group

6.1.1. Email Classification Results Using the Filter Method (Job Category)

For reproducibility, Table 5 summarizes the hyperparameter configurations used for each classifier in this study. All models were implemented using scikit-learn (version 1.3) with consistent settings across all group-level experiments. The results of email classification using the Filter Method for feature selection, categorized by job group, are summarized in Table 6. The findings indicate that job characteristics have a clear influence on classification performance. Job groups exhibiting more consistent and structured email communication behaviors achieved higher accuracy and F1-scores with a relatively small number of selected features. By contrast, groups with more diverse and heterogeneous email patterns require a larger set of features to attain comparable classification performance.

Table 5. Classifier hyperparameter configurations.

Classifier	Hyperparameter	Value
Random Forest	Number of estimators	100
	Splitting criterion	Gini impurity
	Max features per split	$\sqrt{n_{\text{features}}}$
	Min samples per leaf	1
	Max depth	None (fully grown)
Decision Tree (CART)	Algorithm	CART
	Splitting criterion	Gini impurity
	Max depth	None (fully grown)
	Min samples per leaf	1
Logistic Regression	Solver	lbfgs
	Maximum iterations	1000
	Regularization (C)	1.0

The Legal and Policy group (Group J9) achieved the best performance using only two features (Work experience and Working Hour): accuracy 93.86%, F1-score 92.91%, and precision 94.29%. The Customer Service group (Group J7) also performed well (accuracy 88.67%, F1-score 87.75%) with five features, while the Development and Analysis group (Group J3) attained an accuracy of 88.09% and an F1-score of 87.73% using six features. In contrast, Human Resources (Group J4) and IT (Group J2) showed the lowest performance (accuracy $\approx$ 75–76%) despite using a similar number of features, suggesting more heterogeneous behaviors. Notably, Finance and Accounting (Group J6) achieved 83.90% accuracy with only three highly relevant features, reinforcing that feature relevance matters more than feature count.

Table 6. Best classification models and selected features using the filter method (grouped by job categories).

Group	Model	Acc.	F1	Prec.	Rec.	#Feat.	Selected Features (Importance Order)
J1	Decision Tree	79.13%	79.04%	79.03%	79.13%	5	Working hours, Reply/Forward status, Work experience, Sender status, Email size
J2	Random Forest	75.69%	73.53%	73.49%	75.69%	6	Working hours, Reply/Forward status, Sender status, Work experience, Weekday, Email size
J3	Random Forest	88.09%	87.73%	87.69%	88.09%	6	Reply/Forward status, Working hours, Work experience, Email size, Weekday, Sender status
J4	Decision Tree	75.58%	72.16%	75.96%	75.58%	6	Reply/Forward status, Working hours, Work experience, Sender status, Email size, Weekday
J5	Random Forest	81.76%	80.89%	80.70%	81.76%	2	Working hours, Email size
J6	Decision Tree	83.90%	80.30%	82.45%	83.90%	3	Working hours, Reply/Forward status, Work experience
J7	Random Forest	88.67%	87.75%	88.24%	88.67%	5	Work experience, Reply/Forward status, Email size, Working hours, Sender status
J8	Random Forest	77.23%	76.67%	77.17%	77.23%	6	Reply/Forward status, Working hours, Work experience, Sender status, Email size, Weekday
J9	Random Forest	93.86%	92.91%	94.29%	93.86%	2	Work experience, Working hours
J10	Random Forest	81.93%	81.77%	82.11%	81.93%	6	Working hours, Reply/Forward status, Work experience, Email size, Sender status, Weekday
J11	Random Forest	76.32%	76.30%	76.30%	76.32%	6	Reply/Forward status, Work experience, Working hours, Email size, Sender status, Weekday
J12	Random Forest	83.72%	83.72%	83.72%	83.72%	5	Reply/Forward status, Work experience, Working hours, Weekday, Email size

6.1.2. Email Classification Results Using the Filter Method (Email Volume)

Table 7 reports the results when grouped by email volume (volume). Random Forest handled complex inter-feature relations well, especially in medium- to high-volume groups. Group V3 (1001–2000/day) achieved the best performance (accuracy 80.01%, F1 79.99%) with five behavior-centric features. Group V1 (≤ 500 /day) also performed strongly (accuracy 79.29%) with five features. Group V4 (2001–3000/day) had the lowest accuracy (75.77%), suggesting that a high volume introduces variability that reduces model stability.

Table 7. Best classification models and selected features using the filter method (grouped by email volume).

Group	Model	Acc.	F1	Prec.	Rec.	#Feat.	Selected Features (Importance Order)
V1	Random Forest	79.29%	75.57%	80.67%	79.29%	5	Working hours, Reply/Forward status, Work experience, Sender status, Email size category
V2	Random Forest	76.70%	76.61%	76.80%	76.70%	6	Reply/Forward status, Working hours, Work experience, Sender status, Email size category, Weekday
V3	Random Forest	80.01%	79.99%	80.08%	80.01%	5	Reply/Forward status, Working hours, Work experience, Email size category, Weekday
V4	Random Forest	75.77%	75.31%	75.41%	75.77%	6	Reply/Forward status, Working hours, Work experience, Email size category, Weekday, Sender status
V5	Random Forest	76.32%	76.30%	76.30%	76.32%	6	Reply/Forward status, Work experience, Working hours, Email size category, Sender status, Weekday

6.1.3. Email Classification Results Using the Filter Method (Work Experience)

Table 8 shows the results when grouped by Work Experience. Group Y1 (<4 years) achieved the highest accuracy (80.60%) and recall (80.60%), driven by behavior-rich features. Group Y7 (>30 years) reached 79.49% accuracy with only four well-aligned features, whereas Group Y3 (10–14 years) had the lowest accuracy (68.10%) despite using six features, again underscoring that feature relevance outweighs feature count.

Table 8. Best classification models and selected features using the filter method (grouped by work experience).

Group	Model	Acc.	F1	Prec.	Rec.	#Feat.	Selected Features (Importance Order)
Y1	Random Forest	80.60%	75.35%	77.19%	80.60%	6	Working hours, Reply/Forward status, Work experience, Email size category, Weekday, Sender status
Y2	Random Forest	75.22%	72.72%	73.30%	75.22%	6	Reply/Forward status, Working hours, Sender status, Work experience, Weekday, Email size category
Y3	Random Forest	68.10%	66.77%	67.51%	68.10%	6	Working hours, Reply/Forward status, Sender status, Email size category, Weekday, Work experience
Y4	Random Forest	73.98%	73.36%	73.64%	73.98%	6	Reply/Forward status, Working hours, Email size category, Sender status, Weekday, Work experience
Y5	Random Forest	71.11%	68.49%	70.54%	71.11%	6	Reply/Forward status, Working hours, Email size category, Weekday, Work experience, Sender status
Y6	Random Forest	70.21%	68.33%	74.07%	70.21%	6	Reply/Forward status, Sender status, Working hours, Email size category, Work experience, Weekday
Y7	Random Forest	79.49%	73.85%	79.82%	79.49%	4	Reply/Forward status, Working hours, Email size category, Sender status

6.1.4. Email Classification Results Using the Wrapper Method (Job Category)

Wrapper-based results when grouped by job type (Table 9) show that the Decision Tree often excels with compact, well-defined features. The Legal and Policy group (Group J9) again topped all (accuracy 95.33%, F1 94.81%) with only two features. Groups J3 and J7 also performed well, with relatively few features. IT (Group J2) and HR (Group J4) lagged, indicating more diverse behaviors.

Table 9. Best classification models and selected features using the wrapper method (grouped by job categories).

Group	Model	Acc.	F1	Prec.	Rec.	#Feat.	Selected Features (Importance Order)
J1	Random Forest	79.34%	79.35%	79.38%	79.34%	6	Work experience, Reply/Forward status, Sender status, Work experience group, Email size category, Working hours
J2	Random Forest	74.71%	71.42%	71.87%	74.71%	5	Sender status, Weekday, Email size category, Working hours, Work experience group
J3	Decision Tree	88.66%	88.19%	88.17%	88.66%	7	Work experience, Email size category, Reply/Forward status, Working hours, Sender status, Work experience group, Weekday
J4	Decision Tree	76.84%	73.13%	78.72%	76.84%	5	Work experience, Sender status, Weekday, Reply/Forward status, Work experience group
J5	Random Forest	87.32%	86.65%	86.74%	87.32%	2	Weekday, Email size category
J6	Random Forest	84.41%	79.99%	85.08%	84.41%	4	Work experience, Sender status, Weekday, Work experience group
J7	Decision Tree	88.80%	88.37%	88.27%	88.80%	4	Work experience, Reply/Forward status, Sender status, Work experience group
J8	Decision Tree	77.25%	76.70%	77.20%	77.25%	4	Work experience, Working hours, Sender status, Reply/Forward status
J9	Decision Tree	95.33%	94.81%	95.58%	95.33%	2	Work experience group, Working hours
J10	Decision Tree	81.36%	81.15%	81.61%	81.36%	5	Work experience, Working hours, Reply/Forward status, Sender status, Email size category
J11	Decision Tree	76.45%	76.47%	76.51%	76.45%	6	Work experience, Sender status, Reply/Forward status, Weekday, Working hours, Email size category
J12	Decision Tree	83.31%	83.30%	83.31%	83.31%	6	Work experience, Working hours, Reply/Forward status, Weekday, Sender status, Email size category

6.1.5. Email Classification Results Using the Wrapper Method (Email Volume)

Table 10 shows a clear trend: performance depends on both the number and relevance of the features. Group V2 (501–1000/day) had the best overall accuracy (82.80%), while Group V5 (>3000/day) had the lowest accuracy (76.64%), reflecting challenges from extreme volume variability. Group V1 (≤ 500 /day) achieved a solid 80.21% with only two features using a Decision Tree.

Table 10. Best classification models and selected features using the wrapper method (grouped by email volume).

Group	Model	Acc.	F1	Prec.	Rec.	#Feat.	Selected Features (Importance Order)
V1	Decision Tree	80.21%	76.15%	82.88%	80.21%	2	Work experience, Sender status
V2	Random Forest	82.80%	82.65%	83.32%	82.80%	6	Job category, Work experience, Sender status, Reply /Forward status, Working hours, Work experience group
V3	Random Forest	82.03%	82.00%	82.14%	82.03%	7	Work experience, Job category, Working hours, Reply /Forward status, Sender status, Work experience group, Weekday
V4	Random Forest	78.11%	77.83%	77.84%	78.11%	8	Work experience, Job category, Sender status, Reply /Forward status, Working hours, Email size category, Work experience group, Weekday
V5	Random Forest	76.64%	76.63%	76.64%	76.64%	6	Work experience, Sender status, Job category, Weekday, Email size category, Reply /Forward status

6.1.6. Email Classification Results Using the Wrapper Method (Work Experience)

In Table 11, Group Y1 (<4 years) attained the highest overall performance (Accuracy 82.47%) with eight multidimensional features. Group Y7 (>30 years) performed well with only five features. Mid-range tenure groups were moderate, while Groups Y5 and Y6 performed at a lower level, reaffirming that the relevance and representativeness of features are more critical than quantity.

Table 11. Best classification models and selected features using the wrapper method (grouped by work experience).

Group	Model	Acc.	F1	Prec.	Rec.	#Feat.	Selected Features (Importance Order)
Y1	Random Forest	82.47%	80.35%	80.39%	82.47%	8	Sender status, Working hours, Job category, Work experience, Reply /Forward status, Weekday, Email size category, Email volume group
Y2	Random Forest	79.56%	77.37%	79.01%	79.56%	5	Job category, Sender status, Work experience, Working hours, Reply /Forward status
Y3	Random Forest	76.44%	75.83%	76.40%	76.44%	8	Job category, Working hours, Work experience, Sender status, Reply /Forward status, Email volume group, Email size category, Weekday
Y4	Decision Tree	80.26%	80.15%	80.13%	80.26%	6	Job category, Work experience, Reply /Forward status, Sender status, Working hours, Email size category
Y5	Decision Tree	77.59%	76.71%	77.30%	77.59%	8	Sender status, Job category, Work experience, Working hours, Reply /Forward status, Email size category, Weekday, Email volume group
Y6	Decision Tree	75.76%	75.55%	76.09%	75.76%	6	Sender status, Job category, Work experience, Reply /Forward status, Weekday, Email size category
Y7	Random Forest	79.73%	74.19%	80.38%	79.73%	5	Reply /Forward status, Sender status, Email size category, Working hours, Work experience

6.2. Performance of ASPEC

The performance of the ASPEC model was evaluated using the held-out test set of 126,913 emails (30% of the 423,043 emails from the first two months. The remaining 202,026 emails from the third month were reserved exclusively for incremental learning evaluation.) It was compared with the baseline model derived from the experiments in Table 1. The overall comparison results are summarized in Table 12.

Table 12. Comparison of classification performance between the baseline model and ASPEC.

Model	Accuracy	F1-Score	Precision	Recall
Baseline	73.50%	73.37%	71.37%	75.47%
ASPEC	77.79%	76.81%	76.12%	76.42%

The experimental results indicated that the ASPEC model clearly outperformed the baseline model across all performance metrics. Specifically, ASPEC achieved an accuracy of 77.79%, which is substantially higher than the baseline accuracy of 73.50%, where the

Decision Tree was identified as the best-performing algorithm among the baseline models. In addition, the F1-score of ASPEC improved to 76.81%, reflecting a more effective balance between precision and recall.

A more detailed analysis showed that ASPEC maintained a well-balanced trade-off between precision and recall, achieving a precision of 76.12% and a recall of 76.42%. These results demonstrate the model's enhanced capability to accurately identify work-related emails while effectively reducing misclassification rates, particularly false positives, compared with the baseline approach. Overall, the findings confirm the effectiveness of the semi-personalized email classification strategy employed by ASPEC in real-world organizational environments.

To assess the statistical significance of the observed accuracy improvement, 95% confidence intervals were computed for both models using the Wilson score method on the held-out test set derived from the 70:30 split of the first two months' data ($n = 126,913$). The baseline Decision Tree achieved an accuracy of 73.50% with a 95% confidence interval of [73.26%, 73.74%], whereas ASPEC achieved 77.79% with a confidence interval of [77.56%, 78.02%]. The absence of overlap between the confidence intervals suggests that the observed performance improvement is unlikely to be attributable to random sampling variability and is statistically meaningful.

In practical terms, even a moderate improvement in classification accuracy can substantially reduce the volume of incorrectly prioritized emails in large-scale organizational environments. Given the scale of the dataset and the metadata-only deployment constraint, the observed improvement is considered practically meaningful for enterprise email management applications.

6.3. Performance of ASPEC After Incremental Learning

The Incremental Learning process was designed to enable the system to continuously update user-related information in accordance with the evolving data patterns in real-world environments. The primary objective of this process was to enhance the learning capability of the model, ensuring that it remained accurate over time. In this study, 202,026 email records collected over a one-month period were used. The dataset of 202,026 emails was evaluated using a prequential (test-then-learn) protocol, in which each instance was first classified by the current model before being used to update it. This approach ensures that every instance contributes to both evaluation and incremental model refinement without requiring a separate train-test split.

Table 13 demonstrates that applying the Incremental Method does not consistently improve performance across all user groups. In contrast, several cases exhibited a noticeable decline in accuracy compared to the previous performance results, particularly for groups with complex data structures or imbalanced email distributions. For example, the employee group with 20–25 years of work experience (Y5) experienced a substantial decrease in accuracy from 77.59% to 62.15%, representing a reduction of 15.44%. Similarly, the group receiving 501–1000 emails per day (V2) showed a significant decline in accuracy from 82.80% to 71.64%, corresponding to a difference of 11.16%.

These findings indicate that when prior performance results yield higher accuracy, it is advisable to retain the previous model (M_{k-1}) to preserve the reliability and performance standards of the system in real-world deployment. Nevertheless, in some cases, Incremental Learning can maintain performance levels close to those of the previous model. For instance, the Customer Service group (J7) exhibited only a minor reduction in accuracy of 1.09%, decreasing from 88.80% to 87.71%. Likewise, employees with more than 30 years of work experience (Y7) showed almost no change in performance, with accuracy decreasing

marginally from 79.73% to 79.51%, or only 0.22%. These results suggest that the model can be adapted with a high degree of stability for certain user groups.

Moreover, some groups continued to demonstrate strong performance even after Incremental Learning was applied. In particular, the Legal and Policy staff group (J9) maintained the highest accuracy among all groups. Although its Accuracy decreased from 95.33% to 92.57%, its performance remained notably superior and reliable for practical organizational use.

Table 13. Comparison of model accuracy before and after incremental learning across all user groups.

Group	Model (M_{k-1})	Acc. (%)	Model (M_k)	Acc. (%)	Diff (%)
Job Category Groups					
J1	RF	79.34	LR	75.97	−3.37
J2	RF	75.69	DT	73.83	−1.86
J3	DT	88.66	DT	83.48	−5.18
J4	DT	76.84	LR	73.50	−3.34
J5	RF	87.32	LR	81.25	−6.07
J6	DT	83.90	LR	80.47	−3.43
J7	DT	88.80	LR	87.71	−1.09
J8	DT	77.25	LR	71.87	−5.38
J9	DT	95.33	LR	92.57	−2.76
J10	RF	81.93	LR	76.42	−5.51
J11	DT	76.45	LR	69.72	−6.73
J12	RF	83.72	LR	78.38	−5.34
Volume Groups					
V1	DT	80.21	LR	76.52	−3.69
V2	RF	82.80	LR	71.64	−11.16
V3	RF	82.03	LR	75.13	−6.90
V4	RF	78.11	LR	71.69	−6.42
V5	RF	76.64	LR	69.72	−6.92
Work Experience Groups					
Y1	RF	82.47	LR	79.02	−3.45
Y2	RF	79.56	LR	72.69	−6.87
Y3	RF	76.44	LR	68.02	−8.42
Y4	DT	80.26	LR	75.39	−4.87
Y5	DT	77.59	LR	62.15	−15.44
Y6	DT	75.76	LR	67.86	−7.90
Y7	RF	79.73	LR	79.51	−0.22

Note: RF = Random Forest, DT = Decision Tree, LR = Logistic Regression.

Overall, the evaluation of the Incremental Method indicates that although this approach enables the model to accommodate newly arriving data and adapt to evolving user behaviors, its performance may fluctuate depending on the characteristics of the data introduced in each update cycle. Groups with high data diversity or rapidly changing email usage patterns, such as customer service roles and users receiving 501–1000 emails per week, tended to experience more pronounced declines in accuracy after updates. By contrast, groups with more stable email usage patterns, such as employees with over 30 years of experience, are better able to maintain consistent performance despite the incorporation of new data.

These findings suggest that Incremental Learning may be particularly beneficial for users with consistent communication behaviors and relatively stable email patterns. Consequently, model updates should be performed selectively at the group level rather than uniformly across all user groups. This selective updating strategy can help mitigate unnecessary instability while preserving high performance in groups that have already

performed well. These observations are consistent with the findings of Elwell and Polikar [23], who noted that incremental models are sensitive to non-stationary data distributions and concept drift, leading to variability in accuracy depending on the characteristics of the newly observed data. Concept drift refers to changes in the underlying statistical properties of the data distribution over time—formally, a shift in the joint distribution $P(X, y)$ between consecutive model update cycles [23]. In non-stationary environments, where the email distribution of a user group evolves due to organizational changes, seasonal patterns, or shifts in work context, a model trained on historical data may no longer accurately represent the current distribution. The champion-challenger safeguard in Equation (7) mitigates this risk by retaining the previous model $M^{(g,k-1)}$ when the updated model $M^{(g,k)}$ fails to improve classification accuracy, thereby preventing performance degradation caused by abrupt concept drift. This behavior is consistent with the accuracy declines observed in Table 13 for groups such as Y5 (−15.44%) and V2 (−11.16%), which exhibit high behavioral variability and are more susceptible to non-stationary data distributions.

It is important to clarify the intended role of the incremental learning component within ASPEC. The primary objective of incremental learning in this framework is not solely to maximize accuracy at each update cycle, but rather to maintain the system's adaptability to evolving email patterns over time without requiring full model retraining—thereby reducing computational overhead in long-term deployment. The champion-challenger mechanism defined in Equation (7) provides an explicit safeguard: the updated model replaces the previous one only when its accuracy is equal to or higher than that of the predecessor, ensuring that performance cannot degrade below the previously established level through automated model acceptance. The observed accuracy declines in Table 13 reflect cases where the new model (M_k) was outperformed by the previous model (M_{k-1}), meaning the system would automatically retain M_{k-1} in a real deployment scenario. These results, therefore, do not represent actual deployed performance degradation but rather illustrate the distribution of update outcomes across groups. Furthermore, groups with stable communication patterns—such as Customer Service (J7, −1.09%) and employees with over 30 years of experience (Y7, −0.22%)—demonstrated near-negligible performance changes, supporting the viability of selective group-level incremental updating as the recommended deployment strategy.

To provide clearer guidance on when incremental learning is beneficial, the user groups in Table 13 can be categorized into three performance tiers based on the magnitude of accuracy change after updating. Groups with $|\Delta| < 2\%$ (stable tier)—including J7 (−1.09%), Y7 (−0.22%), J2 (−1.86%), and J9 (−2.76%)—are well-suited for regular incremental updates, as their communication patterns are sufficiently stable to accommodate new data without significant performance degradation. Groups with $2\% \leq |\Delta| < 6\%$ (moderate tier) may benefit from selective updating with careful monitoring. Groups with $|\Delta| \geq 6\%$ (unstable tier)—including Y5 (−15.44%), V2 (−11.16%), and Y3 (−8.42%)—exhibit high behavioral variability and should retain the previous model (M_{k-1}) unless a sustained improvement trend is observed over multiple update cycles. This tiered strategy offers a practical deployment guideline for organizations seeking to balance adaptability with classification stability.

7. Conclusions

The proposed ASPEC model has several notable strengths. Its semi-personalized architecture enables adaptation to the diverse characteristics of different user groups, allowing the model to better capture real communication behaviors within an organization, thereby enhancing employees' productivity and reducing work-related stress over the long

term. Additionally, its behavior-driven feature selection approach enhances interpretability and maintains high performance, even with a minimal set of features. The model further benefits from its incremental learning capability, which supports continuous updates and prevents performance degradation as user behaviors or email distributions evolve in dynamic organizational environments. Moreover, the integration of incremental learning allows the model to be retrained using newly emerging data, thereby increasing its flexibility and maintaining alignment with real operational contexts—reinforcing the model’s robustness against data drifts and sudden changes in data distribution. This continuous learning mechanism ensures that the email classification system remains accurate, adaptive, and effective in dynamic, real-world organizational environments.

In terms of computational cost-effectiveness, ASPEC is designed to minimize infrastructure requirements by relying exclusively on classical ML classifiers that can be trained and deployed without GPU resources or large-scale computing infrastructure. The group-level model training is performed offline as a one-time process, and real-time inference requires only a single model lookup per employee profile. At an organizational scale, the observed accuracy improvement of 4.29 percentage points translates to approximately 8667 additional correctly classified emails per month on the evaluation dataset of 202,026 emails—representing a meaningful operational benefit relative to the system’s low computational overhead.

This study has several limitations that indicate directions for future work. First, the absence of direct comparison with transformer-based or deep learning models is acknowledged as a limitation. ASPEC operates exclusively on email metadata without accessing body content—a constraint that prevents the direct application of content-dependent models such as BERT. Extending the framework to incorporate lightweight transformer-based subject line encoders (e.g., DistilBERT applied to translated subject lines) represents a promising direction for future work, particularly for organizations where richer textual features can be made available under appropriate privacy safeguards. Second, ASPEC is best suited for small to medium-sized organizations where computational resources are limited and data privacy is paramount. For very large enterprises, the current model selection mechanism based on three group dimensions may require extension to accommodate greater user diversity. Future work could explore the integration of AI agents or conversational interfaces (e.g., chatbots) to enable dynamic user feedback collection, which could further refine group-level models and enhance personalization beyond the current semi-personalized architecture. Third, the evaluation is conducted within a single organization over a two-month period. While the dataset is large-scale (625,069 emails across three months, 9788 employees), the generalizability of ASPEC to organizations in different sectors, of different sizes, or with different communication cultures remains to be validated. Future work should extend the evaluation to multiple organizations and longer time periods to assess the robustness of the group-level segmentation strategy and the incremental learning mechanism under greater organizational diversity and temporal concept drift.

Author Contributions: Conceptualization, W.K. and N.W.; methodology, W.K.; software, W.K.; validation, W.K. and N.W.; formal analysis, W.K.; investigation, W.K.; data curation, W.K.; writing—original draft preparation, W.K.; writing—review and editing, N.W.; visualization, W.K.; supervision, N.W.; project administration, N.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are not publicly available due to organizational confidentiality constraints and compliance with Thailand’s Personal Data Protection Act (PDPA B.E. 2562). Researchers seeking access to anonymized subsets of the data for replication purposes may contact the corresponding author with a formal data access request, subject to institutional review and organizational approval. All implementation details, including classifier hyperparameters, feature selection configurations, and preprocessing procedures, are fully documented in the revised manuscript to support reproducibility.

Acknowledgments: ChatGPT (GPT-4o) and Gemini 1.5 Pro were used for language editing, including translation and rephrasing to improve clarity and readability. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASPEC	Adaptive Semi-Personalized Email Classification
ML	Machine Learning
DL	Deep Learning
NB	Naïve Bayes
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
TF-IDF	Term Frequency–Inverse Document Frequency
NLTK	Natural Language Toolkit
PDPA	Personal Data Protection Act
GRU	Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
GWO	Gray Wolf Optimizer
SGNN	Semantic Graph Neural Network
USDM	Universal Spam Detection Model
LLM	Large Language Model
CART	Classification and Regression Tree
SVO	Subject–Verb–Object
HR	Human Resources
PII	Personally Identifiable Information
RFC	Request for Comments
SPF	Sender Policy Framework
DKIM	DomainKeys Identified Mail
TCR	Total Cost Ratio
BFA	Binary Firefly Algorithm
MLP	Multilayer Perceptron
AIC	Akaike Information Criterion
DSPF	Dynamic Sender Policy Framework
BIC	Bayesian Information Criterion
SMOTE	Synthetic Minority Oversampling Technique

IP	Internet Protocol
NN	Neural Network
AIS	Artificial Immune System

References

1. Statista. Number of Emails Sent and Received Per Day Worldwide from 2017 to 2025. Available online: <https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/> (accessed on 1 October 2025).
2. Google Workspace Center. Understanding Gmail Categories and Tabs. Available online: <https://support.google.com/mail/answer/3094499> (accessed on 1 October 2025).
3. Resnick, P. Internet Message Format (RFC 5322). *Internet Engineering Task Force (IETF)*. 2008. Available online: <https://www.rfc-editor.org/rfc/rfc5322> (accessed on 1 October 2025).
4. Wong, M.; Schlitt, W. Sender Policy Framework (SPF) for Authorizing Use of Domains in E-Mail, Version 1 (RFC 4408). *Internet Engineering Task Force (IETF)*. 2006. Available online: <https://www.rfc-editor.org/rfc/rfc4408> (accessed on 1 October 2025).
5. Allman, E.; Callas, J.; Delany, M.; Libbey, M.; Fenton, J.; Thomas, M. DomainKeys Identified Mail (DKIM) Signatures (RFC 4871). *Internet Engineering Task Force (IETF)*. 2007. Available online: <https://www.rfc-editor.org/rfc/rfc4871> (accessed on 1 October 2025).
6. Sheu, J.-J.; Chu, K.-T.; Li, N.-F.; Lee, C.-C. An efficient incremental learning mechanism for tracking concept drift in spam filtering. *PLoS ONE* **2017**, *12*, e0171518. [[CrossRef](#)] [[PubMed](#)]
7. Martin, S.; Sewani, A.; Nelson, B.; Chen, K.; Joseph, A.D. Analyzing behavioral features for email classification. In Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS 2005), 21–22 July 2005, Stanford, CA, USA; pp. 1–10.
8. Aslantaş, V.; Khorsheed, F.H.; Ahmed, B. Wrapper feature selection approach based on binary firefly algorithm for spam e-mail filtering. *J. Soft Comput. Data Min.* **2020**, *1*, 44–52.
9. Nasreen, G.; Khan, M.M.; Younus, M.; Zafar, B.; Hanif, M.K. Email spam detection by deep learning models using novel feature selection technique and BERT. *Egypt. Inform. J.* **2024**, *26*, 100473. [[CrossRef](#)]
10. Cormack, G. Email spam filtering: A systematic review. *Found. Trends Inf. Retr.* **2006**, *1*, 335–455. [[CrossRef](#)]
11. Sahami, M.; Dumais, S.; Heckerman, D.; Horvitz, E. A Bayesian approach to filtering junk e-mail. In Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, 26 July 1998, Madison, WI, USA; pp. 55–62.
12. Nguyen, T.A.; Tran, Q.A.; Nguyen, X.T. Spam filter based on dynamic sender policy framework. In Proceedings of the 2nd International Conference on Knowledge and Systems Engineering (KSE 2010), 7–9 October 2010, Hanoi, Vietnam; pp. 224–228.
13. Ajani, T.; Ferrante, T. Cyber-analytics: An examination of machine learning algorithms for spam filtering. *Issues Inf. Syst.* **2024**, *25*, 203–213.
14. Awad, W.A.; Elseuofi, S.M. Machine learning methods for spam e-mail classification. *Int. J. Comput. Sci. Inf. Technol.* **2011**, *3*, 173–184. [[CrossRef](#)]
15. Androutsopoulos, I.; Koutsias, J.; Chandrinou, K.V.; Paliouras, G.; Spyropoulos, C.D. An evaluation of naive bayesian anti-spam filtering. In *Proceedings of the Workshop on Machine Learning in the New Information Age*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 9–17.
16. Christina, V.; Karpagavalli, S.; Suganya, G. Email spam filtering using supervised machine learning techniques. *Int. J. Comput. Sci. Eng.* **2010**, *2*, 3126–3129.
17. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2–7 June 2019, Minneapolis, MN, USA; pp. 4171–4186.
18. AbdulNabi, I.; Yaseen, Q. Spam email detection using deep learning techniques. *Procedia Comput. Sci.* **2021**, *184*, 853–858. [[CrossRef](#)]
19. Muhammad, J.M.; Hasan, A.B.; Farrukh, M. Classification and prediction of spam emails based on AI enabling models using deep and machine learning techniques. In *Proceedings of the 2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
20. Pan, W.; Li, J.; Gao, L.; Yue, L.; Yang, Y.; Deng, L.; Deng, C. Semantic graph neural network: A conversion from spam email classification to graph classification. *Sci. Program.* **2022**, *2022*, 8. [[CrossRef](#)]
21. Halder, T.; Srizon, A.Y.; Esha, N.T.; Hasan, S.M.M.; Faruk, M.F.; Hossain, M.R. Enhancing email safety: Harnessing ML, DL, and LLM models for spam detection. In *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
22. Tida, V.S.; Hsu, S. Universal spam detection using transfer learning of BERT model. In Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS), 4–7 January 2022, Virtual; pp. 7849–7858.

23. Elwell, R.; Polikar, R. Incremental learning of concept drift in nonstationary environments. *IEEE Trans. Neural Netw.* **2011**, *22*, 1517–1531. [[CrossRef](#)] [[PubMed](#)]
24. Mang, L.X.; Jung, H.; Youn, H.Y.; Kim, U.-M. An incremental learning based framework for image spam filtering. *Int. J. Comput. Sci. Eng. Appl.* **2014**, *4*, 25–40. [[CrossRef](#)]
25. Kang, C.; Shang, H.; Langlois, J.-M. Classifying emails into human vs machine category. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), 22 February–1 March 2022, Online; pp. 7069–7077.
26. Shu, K.; Mukherjee, S.; Zheng, G.; Awadallah, A.H.; Shokouhi, M.; Dumais, S. Learning with weak supervision for email intent detection. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 25–30 July 2020, Virtual; pp. 1051–1060.
27. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [[CrossRef](#)]
28. Bird, S.; Loper, E. NLTK: The natural language toolkit. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, 21–26 July 2004, Barcelona, Spain; pp. 214–217.
29. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **1972**, *28*, 11–21. [[CrossRef](#)]
30. Omotehinwa, T.O.; Oyewola, D.O. Hyperparameter optimization of ensemble models for spam email detection. *Appl. Sci.* **2023**, *13*, 1971. [[CrossRef](#)]
31. Yusupov, K.; Islam, M.R.; Muminov, I.; Sahlabadi, M.; Yim, K. Comparative analysis of machine learning and deep learning models for email spam classification using TF-IDF and word embedding techniques. In *Advances on Broad-Band Wireless Computing, Communication and Applications (BWCCA 2024)*; Barolli, L., Ed.; Lecture Notes on Data Engineering and Communications Technologies; Springer: Cham, Switzerland, 2025; Volume 231, pp. 114–122. [[CrossRef](#)]
32. Kontsewaya, Y.; Antonov, E.; Artamonov, A. Evaluating the effectiveness of machine learning methods for spam detection. *Procedia Comput. Sci.* **2021**, *190*, 479–486. [[CrossRef](#)]
33. Forman, G. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* **2003**, *3*, 1289–1305.
34. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.