Practical Spam Email Detecting Algorithms: a Comparison Study with Continuous Training and Lightweight Personalization

Zixuan Sun
Siebel School of Computing and Data Science, UIUC
Champaign, Illinois, USA
zixuans8@illinois.edu

Junjie Ao
Grainger School of Engineering, UIUC
Champaign, Illinois, USA
junjiea2@illinois.edu

Yueqiang Wu
Grainger School of Engineering, UIUC
Champaign, Illinois, USA
wu147@illinois.edu

Abstract

Spam email detection and filtering remains an open problem due to the evolving strategies of spammers and the instability of existing models in real-world settings. To solve this problem, classical machine learning approaches provide lightweight yet competitive baselines [3], while modern Neural-Network approaches raise unavoidable higher computational costs for both training and deployment [10]. In this project, we aim to systematically compare the performance of these two families of methods across representative public datasets. Through controlled, realistic experiments and incremental improvements, we seek to gather actionable insights into whether inexpensive modifications to baseline models can enhance real-world model performance in a meaningful way.

Keywords: Spam Email Detection, Classification, Continuous Training

ACM Reference Format:
Zixuan Sun, Junjie Ao, and Yueqiang Wu. 2018. Practical Spam Email Detecting Algorithms: a Comparison Study with Continuous Training and Lightweight Personalization. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Spam email detection and filtering remains an open problem in the area of NLP. Even though there exists strong baseline models such as Naive Bayes and modern Transformer-based approaches, but real-world applications of these methods still show lack of stability. In addition, attackers, who know how filtering algorithms work, constantly change language, layout, and sending infrastructure to bypass filters. On the other hand, the performance of a pre-built classifier that has been trained often downgraded performance over time due to concept drift [6] [8], and varies in a large range based on personal context [**?**], which often shifts over time as well.

Our proposed project lies at the intersection of classical text classification and modern ML methods. Traditional methods, including but not limited to Naive Bayes, Logistic Regression, and Support Vector Machines, which are usually trained over retrieved features from the text such as TF-IDF features, still remain fast and competitive these days. Modern Transformer-based models, such as DistillBert [9], shows exceptional skills at extracting meaning and information from text [2][1]. However, a deployable transformer-based model requires much higher cost to both train and retrain [4][10].

Therefore, the main purpose of this project is to implement and compare existing baseline detection methods to gain insight into how their performance varies over time and based on contexts. Moreover, we aim to experiment with simple and computationally inexpensive modifications, such as continuous training through reinforcement leaning techniques and lightweight personal adjustment, to improve performance without adding complexity to the models. Our project will adhere to realistic experimental design: handling real email artifacts, adding time-aware feature extraction, and layering personalized updates in a controlled way. We expect the final result to show guidance on whether simple tactics we experiment with are worth adopting for practical spam filters.

2 Proposed Work

This project mainly focus on implementing, testing, and comparing existing models on widely used datasets. We will use the gathered insights to experiment with modifications on these models.

2.1 Selected Models

In this project, we will not attempt to invent new spam filtering algorithms, as this is a short project comparing

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

to actual research projects. Instead, we will focus on well-known baseline models of the following two types:

1. Traditional Models (over TF-IDF features): Naive Bayes, Multivariate Gaussian Classifier, Logistic Regression with a linear MLP layer and Linear SVM.
2. Modern Models (Transformer-based): DistilBERT [9] and RoBERTA [5].

Traditional models will be implemented and trained from scratch, while we will use pre-trained BERT models and fine-tune on our selected datasets.

2.2 Selected Dataset

We propose to use two public spam email datasets that are small enough to run locally yet representative of real email: a preprossed version of the Apache SpamAssassin Public Corpus [7] and a preprossed version of Enron-Spam Data [11]. We will rely on SpamAssassin dataset to test baseline performance, and test on Enron-Spam to compare performance across personal preferences since Enron-Spam contains information on per-user thresholds and incremental updates on a user's history

2.3 Proposed Modifications

Our initial ideas on possible modifications to experiment with includes compare a static model to an incrementally updated model using rolling windows or partial fit during training to test the effect of continuous training on handling concept drift. As for lightweight personalization, we propose to add a per-user incremental head or dynamically adjusting per-user decision thresholds. In addition, some supplementary modules will be introduced to handle non-textual data in the dataset (such as CNN head for images).

2.4 Evaluation Metrics

We propose to evaluate and compare models through various metrics to analyze their performance on our proposed problems:

1. Primary Performance Score: F1 Score, and high Recall score to force high cost of false positives.
2. Generalization Test: train on one dataset and test on the other, and evaluate performance with primary metric.
3. Time-aware Experiment: Split the dataset based on time stamp, train models on earlier data and test them on later data, and evaluate performance with primary metric.
4. Personalization Analysis: per-user scores on Enron dataset, as well as variance across different users.

2.5 General Timeline and Work Distribution

Week 1 - Data cleaning: Preprocess datasets by each group members, while Zixuan will explore more practical datasets (such as Trec 2007 or UCI Spambase) to run addition tests.
Week 2 - Traditional models: Implement and test traditional models seperately by each group member (one model each person, where logistic regression and Multivariate Gaussian classifier will be assigned to Zixuan).
Week 3 - Transformer-based models: Load, fine-tune, and test BERT models (one model each person, where Yueqiang will explore other models that are potentially applicable).
Week 4 - Evaluate Results: Gather results and compare models to gain insights on possible modifications.
Week 5 - Time-aware Experiment: Proceed with Time-aware Experiment on selected models.
Week 6 - Personalization Analysis: Proceed with personalization analysis on selected models.
Week 7 - Modifications: Experiment with proposed modifications (additional features may also be tested).
Week 8 - Conclusion and Report: Gathering results and present analysis.

References

[1] Suhaima Jamal and Hayden Wimmer. 2023. An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach. arXiv:2311.04913 [cs.CL] https://arxiv.org/abs/2311.04913

[2] Suhaima Jamal, Hayden Wimmer, and Iqbal H. Sarker. 2024. An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. *Security and Privacy* 7, 5 (April 2024), 21 pages. doi:10.1002/spy2.402

[3] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). Association for Computational Linguistics, Valencia, Spain, 427–431. https://aclanthology.org/E17-2068/

[4] Zijie Lin, Zikang Liu, and Hanbo Fan. 2025. Improving Phishing Email Detection Performance of Small Large Language Models. arXiv:2505.00034 [cs.CL] https://arxiv.org/abs/2505.00034

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692 (2019). https://api.semanticscholar.org/CorpusID:198953378

[6] José Márcio Martins da Cruz and Gordon V. Cormack. 2009. Using old Spam and Ham Samples to Train Email Filters. In *Sixth Conference on Email and Anti-Spam - CEAS 2009*. Microsoft Research Silicon Valley, Mountain View, CA, United States. https://hal.science/hal-04691528

[7] Ganiyu Olalekan. 2021. Email Classification (SpamAssassin Email Classification Dataset). Kaggle. https://www.kaggle.com/datasets/ganiyuolalekan/spam-assassin-email-classification-dataset Dataset.

[8] David Ruano-Ordás, Florentino Fdez-Riverola, and José R. Méndez. 2018. Concept drift in e-mail datasets: An empirical study with practical implications. *Information Sciences* 428 (2018), 120–135. doi:10.1016/j.ins.2017.10.049

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019). https://api.semanticscholar.org/CorpusID:203626972

[10] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (Apr.

2020), 13693–13696. doi:10.1609/aaai.v34i09.7123

[11] Marcel Wiechmann. 2021. Enron Spam Data. Kaggle. https://www.kaggle.com/datasets/marcelwiechmann/enron-spam-data Dataset.