The Professional Filter: Machine Learning Approach to Work Email Classification

Ayush Gala
agala2@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Akshay Dongare
adongar@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Vidhisha Kamat
vskamat@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Abstract

This paper presents a machine learning-based approach to email classification, focusing on distinguishing work-related emails from personal ones. Using a manually annotated subset of 5,028 samples from the Enron email corpus, we engineered a diverse set of lexical, semantic, and structural features. Semantic features were extracted using TF-IDF scores and GloVe embeddings, while Latent Dirichlet Allocation (LDA) was employed for topic modeling. Four machine learning models—Support Vector Machines (SVM), Random Forest, AdaBoost, and Artificial Neural Networks (ANN)—were trained and evaluated for their classification performance. The methodology emphasizes semantic modeling to improve generalizability beyond dataset-specific characteristics observed in prior studies. Results demonstrate the potential of combining lexical and semantic features for accurate classification, with insights into feature importance and model performance. The study contributes both methodological insights and newly annotated data samples to the field of email classification. The paper also highlights avenues for integrating Large Language Models to further enhance classification accuracy.

Keywords

Email classification, machine learning, natural language processing, enron dataset

ACM Reference Format:
Ayush Gala, Akshay Dongare, and Vidhisha Kamat. 2024. The Professional Filter: Machine Learning Approach to Work Email Classification. In . ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The growing reliance on email communication has introduced a variety of challenges related to information management and classification. Email classification, which includes tasks such as spam filtering, priority assignment, and automated foldering, has become an essential area of research. This study focuses on a specific classification problem: differentiating work-related emails from personal ones. Accurate classification of emails into these categories can enhance productivity and streamline tasks such as archiving, context-aware notifications, and email prioritization.

The motivation for this project stems from observations made during the analysis of the Enron email corpus. According to Jabbari et al., approximately 20% of the 600,000 emails in the corpus are not work-related, which posed significant challenges during the federal investigation of the Enron scandal. Beyond legal investigations, a reliable professional filter has broad applications in improving email-related tasks, as outlined by Klimt et al. By accurately identifying work-related emails, organizations can achieve better email management and analysis, reducing inefficiencies.

Prior work in this area has focused primarily on using social-network-based features extracted from specific datasets. For example, Gilbert et al. and Alkhereyf et al. explored the Enron dataset through network-centric models, which relied heavily on sender-receiver relationships and hierarchical structures. While these approaches achieved notable success within the bounds of specific datasets, their dependency on dataset-specific features limits their generalizability to other contexts. To address this, our project emphasizes semantic features, which offer a more adaptable framework for classification.

1.1 Literature Survey

Jabbari et al. (2006) conducted a large-scale annotation project on the Enron email corpus, classifying emails into "Business" and "Personal" categories1. This work represents one of the earliest attempts to create a labeled dataset for email classification tasks using the Enron corpus. The authors also sub-categorized emails within these main categories, providing a more granular classification scheme. This paper laid the groundwork for future research on automated email classification and highlighted the potential of the Enron corpus for such tasks.

Klimt and Yang (2004) introduced the Enron corpus as a new benchmark dataset for email classification research2. They analyzed its suitability for email folder prediction tasks and provided baseline results using Support Vector Machines (SVMs) under various conditions. The authors explored the effectiveness of using individual email sections (From, To, Subject, and body) as input features, as well as combining them using regression weights. Their work demonstrated the value of the Enron corpus for email classification research and established initial performance benchmarks for future studies.

Alkhereyf and Rambow (2017) presented an empirical study on classifying emails as business or personal using both lexical and social network features3. They trained their models on the Enron corpus and tested them on both Enron and Avocado corpora. The authors compared the performance of Support Vector Machines

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

(SVM) and Extra-Trees classifiers using these features. Their key findings include: Information from email exchange networks improves classification performance. Combining graph features with lexical features enhances performance for both classifiers. The study provides manually annotated sets of the Avocado and Enron email corpora as a supplementary contribution. This work demonstrates the value of incorporating social network information in email classification tasks and provides insights into cross-corpus generalization.

Agarwal et al. (2012) developed a comprehensive gold standard for the Enron organizational hierarchy4. This resource is valuable for researchers working on tasks related to organizational structure and communication patterns within the Enron corpus. The gold standard can be used to evaluate algorithms that attempt to infer organizational hierarchies from email communications or to study how information flows within corporate structures. Mitra and Gilbert (2013) conducted an exploratory study of gossip in the Enron email dataset5. Using natural language processing techniques, they analyzed the prevalence and characteristics of gossip in workplace emails. This study provides empirical evidence for gossip theories originating from anthropology and offers insights into workplace communication dynamics.

Graus et al. (2014) developed a system for recipient recommendation in enterprise email systems using communication graphs and email content6. While not directly focused on classification, this work demonstrates the application of machine learning techniques to email-related tasks in enterprise settings. The authors likely used features derived from both the communication network structure and the content of emails to make recommendations.

Bekkerman et al. conducted benchmark experiments on email categorization using the Enron and SRI corpora8. They likely compared various machine learning algorithms for the task of automatically sorting emails into user-defined folders. This work provides valuable insights into the practical application of email classification techniques and the challenges associated with personalizing such systems to individual users' folder structures.

2 Methodology

The methodology involves a six-stage workflow to classify emails as work-related or non-work-related. It begins with parsing emails from MIME format into a machine-learning-compatible structure, followed by feature engineering to extract attributes like the number of attachments, recipients, hyperlinks, and email length. Over 5000 emails are then labeled manually. Preprocessing steps, including normalization and dataset splitting, prepare the data for training. Finally, multiple models are trained and evaluated using binary classification metrics such as accuracy, precision, recall, and F1-score.

2.1 Novelty

The novelty of this project lies in its focus on semantic and lexical features, which allow the classifier to generalize beyond specific datasets. We employ Latent Dirichlet Allocation (LDA) to identify dominant topics within email bodies and align them with work or personal classifications. Integration of non-textual features, such as the number of recipients, attachments, and email length, providing additional context for classification. This approach differs slightly from prior studies that relied heavily on social network-based features extracted from sender-receiver relationships.

2.2 Rationale

The methodology is expected to perform well due to its comprehensive feature set and the complementary strengths of the chosen models. The choice of algorithms is informed by their complementary capabilities and their demonstrated success in similar tasks, as outlined in the literature review. SVMs have an ability to handle high-dimensional feature spaces effectively, leveraging the semantic richness of email content. Random Forest models rank feature importance and provide interpretable results, aiding in understanding the characteristics of work-related emails. Adaboost models display high precision in classification tasks and robustness to small variations in the data, ensuring consistent performance. ANNs have the ability to capture complex patterns and relationships within the semantic and structural features, providing a versatile solution.

3 Implementation

3.1 Dataset

The Enron Email Corpus is a large database of over 600,000 emails generated by 158 employees of the Enron Corporation between 1999 and 2002, leading up to the company's collapse. The corpus contains approximately 500,000 unique emails after removing duplicates, with the data spanning about 3,500 user-defined folders. Structurally, the corpus is organized into folders corresponding to individual employees. Each email message typically includes metadata such as sender and recipient email addresses, date and time, subject, and body text. The dataset does not include email attachments.

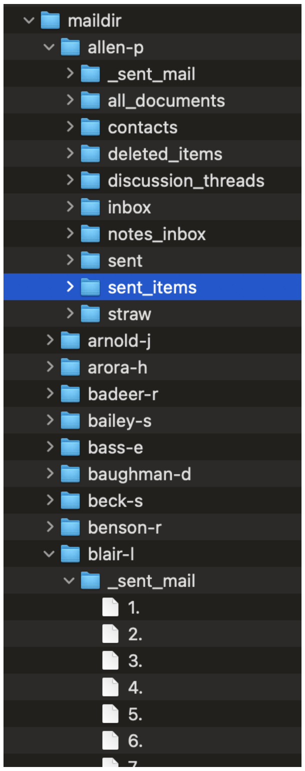
For this study, we manually annotated a total of 5,028 samples into two categories: work-related (1) and personal (0). The manual labeling was aided by the Label Studio tool. Simultaneously, the MIME format of the email was parsed to extract features like subject, domain name, number of recipients, email length, and sent time. We also check if the email has been forwarded or not by checking for subject markers like 'FWD' or 'RE'. The semantic embeddings (using the GloVe 6B model) were computed and were condensed by calculating frequency average of the vector embedding. Lastly, we identified the most dominant topics of each email using Latent Dirichlet allocation and assigned them as an additional feature.

3.2 Hypothesis

The hypothesis for this study is grounded in the premise that work-related and personal emails exhibit distinguishable lexical, semantic, and structural patterns. By leveraging these features, machine learning models can achieve high classification accuracy.

3.3 Workflow Details

The implementation of this project was carried out in three structured phases: 1. Data annotation and feature extraction 2. Feature engineering and 3. Model training and evaluation. Each phase focused on systematically addressing the requirements of the email



```
maildir
  allen-p
    _sent_mail
    all_documents
    contacts
    deleted_items
    discussion_threads
    inbox
    notes_inbox
    sent
    sent_items
    straw
  arnold-j
  arora-h
  badeer-r
  bailey-s
  bass-e
  baughman-d
  beck-s
  benson-r
  blair-l
    _sent_mail
      1.
      2.
      3.
      4.
      5.
      6.
```

Figure 1: Enron corpus folder structure

classification task and ensuring the effective application of machine learning techniques.

3.3.1 Data annotation and feature extraction. :

Emails from the Enron dataset were categorized into work-related and personal classes based on manual inspection. This involved crawling and parsing the folder structure, as outlined by Agarwal et al., to extract relevant emails. Headers and content were carefully reviewed to ensure accurate labeling, creating a high-quality dataset for training machine learning models. Label Studio was employed to facilitate the annotation process, while Python scripts were used for parsing the email structure.

3.3.2 Feature Engineering. :

The second phase focused on feature engineering, wherein lexical, semantic, and structural attributes of the emails were extracted. Lexical features were derived using Term Frequency-Inverse Document Frequency (TF-IDF), capturing the significance of words across the dataset. Semantic embeddings were computed with the GloVe 6B model, leveraging co-occurrence statistics to encode word relationships effectively. Structural features, such as email length, the number of recipients, and the presence of attachments, were programmatically calculated to incorporate contextual information. Topic modeling was conducted using Latent Dirichlet Allocation (LDA) to identify dominant topics within the email content. Finally, preprocessing steps such as scaling and normalization of features ensured uniformity and compatibility across the machine learning models.

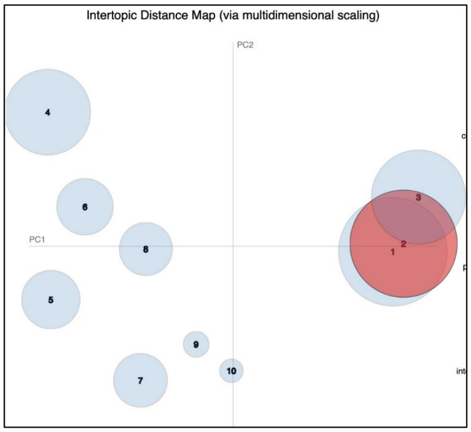


Figure 2: Enron corpus folder structure

3.3.3 Model Training. :

Four distinct models, chosen for their strengths in handling high-dimensional and complex data:

Support Vector Machines (SVM): SVMs were selected for their capability to manage high-dimensional feature spaces effectively. The radial basis function kernel was utilized, with a regularization parameter set to 10 to balance the trade-off between maximizing the margin and minimizing classification errors.

Random Forest: This model was configured with 100 estimators and employed the Gini impurity criterion to measure the quality of splits. Random Forest was chosen for its ability to evaluate feature importance, offering insights into the most significant attributes influencing classification.

AdaBoost: This boosting algorithm was configured with 100 estimators and a learning rate of 0.1. The model's robustness to overfitting and its capability to distinguish fine differences between work-related and personal emails made it an excellent choice for this task. A 5-fold cross-validation approach was used to ensure the reliability of results.

Artificial Neural Networks (ANN): The ANN architecture consisted of two hidden layers, employing ReLU activations for non-linearity and a sigmoid activation in the output layer for binary classification. Early stopping was implemented to prevent overfitting during the 50 training epochs. TensorFlow was used to build and train the neural network.

Various Python libraries, including Scikit-learn, TensorFlow, and NLTK, supported the implementation and evaluation of these models, streamlining the computational processes and ensuring efficiency.

4 Results

The results from the model evaluations indicate that the Random Forest model outperformed the other models, achieving the highest accuracy of 0.9095. The ANN model followed with an accuracy of 0.8559, demonstrating comparable performance to the SVM model, which had an accuracy of 0.8519. The AdaBoost model had the lowest accuracy at 0.8499. Precision and recall metrics further highlight the strengths and weaknesses of the models. The SVM model

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.8519	0.82	0.85	0.83
Random Forest	0.9095	0.90	0.91	0.90
ANN	0.8558	0.83	0.86	0.85
AdaBoost	0.8499	0.85	0.85	0.85

Table 1: Summary of Evaluation metrics

showed high precision for class 1 (0.86) but had a low recall for class 0 (0.15), suggesting it was effective in identifying work-related emails but missed many non-work-related ones. The Random Forest model displayed strong precision and recall for class 1 (0.93 and 0.96, respectively), indicating a balanced classification capability. The ANN model demonstrated reasonable precision and recall, with values of 0.86 and 0.98 for class 1 and 0.68 and 0.17 for class 0. AdaBoost, despite achieving a competitive precision for class 1 (0.85), showed a low recall for class 0 (0.06), leading to a lower F1-score for this class (0.12). These results suggest that the Random Forest model provided the most balanced performance across all metrics, while AdaBoost, despite an overall accuracy close to other models, exhibited challenges in detecting non-work-related emails effectively.

Model	Precision	Recall	F1-Score	Accuracy
SVM	0.63	0.15	0.24	0.8519
Random Forest	0.77	0.62	0.68	0.9095
ANN	0.68	0.17	0.27	0.8559
AdaBoost	0.83	0.06	0.12	0.8499

Table 2: Metrics for Class 0.0 (Personal Emails) for All Models

4.1 Discussion

The results of the project indicate a varied performance across the models tested, highlighting the influence of different algorithmic characteristics on the email classification task. This section provides a detailed analysis of the models, their strengths and weaknesses, and potential limitations, particularly considering the class imbalance present in the dataset.

4.1.1 Support Vector Machine (SVM). : The SVM model achieved an accuracy of 0.8519, with a precision of 0.63 and a recall of 0.15 for class 0.0 (personal emails). The F1-score for this class was 0.24, indicating that while the model maintained overall accuracy, it struggled to effectively identify personal emails. The high precision for class 1.0 (work-related emails), paired with a significantly lower recall for class 0.0, suggests that the model was biased toward correctly classifying work-related emails. This may have been due to the high-class imbalance (82.3% work-related and 17.7% personal emails) present in the data, which could lead the model to prioritize the majority class for higher overall accuracy. The radial basis function kernel used in the SVM did not appear sufficient to address the class imbalance effectively, leading to suboptimal recall for the minority class.

4.1.2 Random Forest. : The Random Forest model demonstrated better overall performance, with an accuracy of 0.9095 and an F1-score of 0.68 for class 0.0. Precision for class 0.0 was 0.77, and recall was 0.62, which were the highest among the models tested. The relatively good recall indicates that the Random Forest model was more capable of capturing personal emails than the SVM, likely due to its ensemble nature that can manage feature importance and learn complex relationships within the data. The feature importance analysis revealed that features like Feature 4 and Feature 1 had the most significant impact on model performance, suggesting that these attributes provided strong signals for distinguishing between work-related and personal emails. However, despite this improved performance, the model still fell short of achieving balanced performance across classes, with recall for the minority class lower than desired.

4.1.3 Artificial Neural Network (ANN). : The ANN model achieved an accuracy of 0.8559 and an F1-score of 0.27 for class 0.0. The precision for this class was 0.68, while recall was 0.17. These results indicate that while the ANN was able to learn some meaningful features from the data, it was not effective at distinguishing personal emails from work-related ones. The lower recall and F1-score suggest that the model may have suffered from overfitting due to the high-dimensional nature of the feature set and insufficient regularization, despite using early stopping as a preventive measure. The model architecture with two hidden layers and ReLU activation functions was capable of capturing non-linear patterns, but the imbalance issue likely hindered its ability to generalize to class 0.0.

4.1.4 AdaBoost. : AdaBoost achieved an accuracy of 0.8499 and an F1-score of 0.12 for class 0.0. This model exhibited a high precision of 0.83 for class 0.0 but a very low recall of 0.06, reflecting an inability to identify most personal emails while maintaining accuracy through class 1.0. The high precision suggests that when the model did classify an email as personal, it was correct, but the recall indicates that most personal emails were missed. The imbalance in the dataset had a notable effect, as the AdaBoost algorithm adjusted weights on misclassified instances in an attempt to improve performance but struggled to generalize to the minority class effectively.

5 Conclusion

Among the models tested, Random Forest demonstrated the most balanced performance, achieving the highest precision and recall for class 0.0. This suggests that Random Forest's ensemble nature, with the ability to learn from various decision trees and evaluate feature importance, provided an advantage in mitigating the effects of class imbalance compared to other models. However, even with Random Forest, the recall was not sufficient to classify personal emails effectively, indicating a need for further enhancement.

SVM, despite its high overall accuracy, struggled with low recall for class 0.0, showcasing its limitations in handling highly imbalanced datasets without additional techniques like class weights or kernel tuning. The ANN model showed that deep learning architectures could capture complex patterns but required further optimization to handle class imbalance effectively. AdaBoost, while achieving a high precision for class 0.0, failed to maintain sufficient recall, reinforcing the challenges of boosting algorithms under imbalanced conditions.

The class distribution, with 82.3% work-related and 17.7% personal emails, had significant implications for all models. The high class imbalance led to biased performance, where models were inclined to favor the majority class to maximize overall accuracy. This bias was observed across all models, as evidenced by the varying recall scores for class 0.0. Techniques such as class weighting, oversampling of the minority class, or advanced resampling strategies could have potentially mitigated this imbalance and improved the recall for class 0.0.

The implications of this class imbalance are critical in practical applications. For example, an email classifier that fails to identify personal emails effectively can lead to user dissatisfaction, missed context-aware notifications, and poor email management. The recall for class 0.0 was particularly low across models, indicating a potential area for improvement in handling class imbalance, such as integrating synthetic data generation or alternative loss functions to account for minority class under-representation.

6 Future improvements

To address the issues observed, several improvements could be made:

Resampling Techniques: Implementing oversampling methods such as SMOTE or undersampling for the majority class could potentially balance the class distribution in training data.

Class Weight Adjustments: Applying class weights during training to penalize misclassifications of the minority class more heavily could improve recall.

Alternative Algorithms: Exploring algorithms that are specifically designed to handle imbalanced data, such as ensemble methods tailored for class imbalance or cost-sensitive learning approaches, might yield better results.

Feature Engineering: Enhancing feature extraction by including more domain-specific features or advanced natural language processing (NLP) techniques, such as BERT embeddings, could improve model generalization. Hyperparameter Tuning: Extending hyperparameter tuning, particularly for models like SVM and ANN, could help find optimal configurations that improve performance for minority class detection.

References

Sanaz Jabbari, Ben Allison, David Guthrie, and Louise Guthrie. 2006. Towards the Orwellian nightmare: separation of business and personal emails. In Pro- ceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, pages 407–411.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In Machine learning: ECML 2004, Springer, pages 217–226.

Work Hard, Play Hard: Email Classification on the Avocado and Enron Corpora" by Alkhereyf and Rambow. This paper presents an empirical study on classifying emails as business or personal using lexical and social network features, training on the Enron corpus and testing on both Enron and Avocado corpora.

Apoorv Agarwal, Adinoyi Omuya, Aaron Harnly, and Owen Rambow. 2012. A comprehensive gold stan- dard for the enron organizational hierarchy. In Pro- ceedings of the 50th Annual Meeting of the Associ- ation for Computational Linguistics: Short Papers- Volume 2. Association for Computational Linguis- tics, pages 161–165.

Tanushree Mitra and Eric Gilbert. 2013. Analyzing gossip in workplace email. ACM SIGWEB Newslet- ter Winter 5.

David Graus, David Van Dijk, Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. 2014. Recip- ient recommendation in enterprises using commu- nication graphs and email content. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, pages 1079–1082.

Maja Pavlovic, Massimo Poesio "The Effectiveness of LLMs as Annotators: A Comparative Overview and Empirical Analysis of Direct Representation" Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024

"Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora" by Bekkerman et al.

A Meeting Attendance

Date	Task
10/20/2024	Conducted project discussion and distributed tasks among team members.
10/21/2024	Reviewed various research papers to gain insights and identify relevant methodologies.
10/23/2024	Implemented folder structure crawling based on the provided descriptions.
10/27/2024	Began manual labeling of samples into "work-related" or "not work-related" categories.
10/28/2024	Continued manual labeling of samples into "work-related" or "not work-related."
10/30/2024	Continued manual labeling of samples into "work-related" or "not work-related."
11/03/2024	Completed manual labeling of samples.
11/04/2024	Finished manual labeling of samples into "work-related" or "not work-related."
11/06/2024	Extracted TF-IDF scores and generated vector embeddings for textual data using the GloVe 6B model.
11/10/2024	Applied Latent Dirichlet Allocation (LDA) to assign dominant topics to the data.
11/11/2024	Condensed vector embeddings through frequency averaging.
11/13/2024	Performed scaling and normalization of features for model input.
11/17/2024	Implemented SVM and Random Forest algorithms for initial classification.
11/18/2024	Explored AdaBoost and Artificial Neural Networks for further model evaluation.
11/20/2024	Held discussions on results, analyzed performance metrics, and identified areas for improvement.
11/24/2024	Started preparation for project presentation.
11/25/2024	Drafted and refined sections of the project report.

Table 3: Work Log for Project Timeline

Our team worked closely together throughout the project, with all three of us attending every meeting. We met regularly, both online and in person at the Hunt Library