The Professional Filter :
Machine Learning Approach to Work Email Detection

Group 6

1. Akshay Dongare (adongar)
2. Ayush Gala (agala2)
3. Vidhisha Kamat (vskamat)

https://github.ncsu.edu/adongar/G6_ALDA_Project

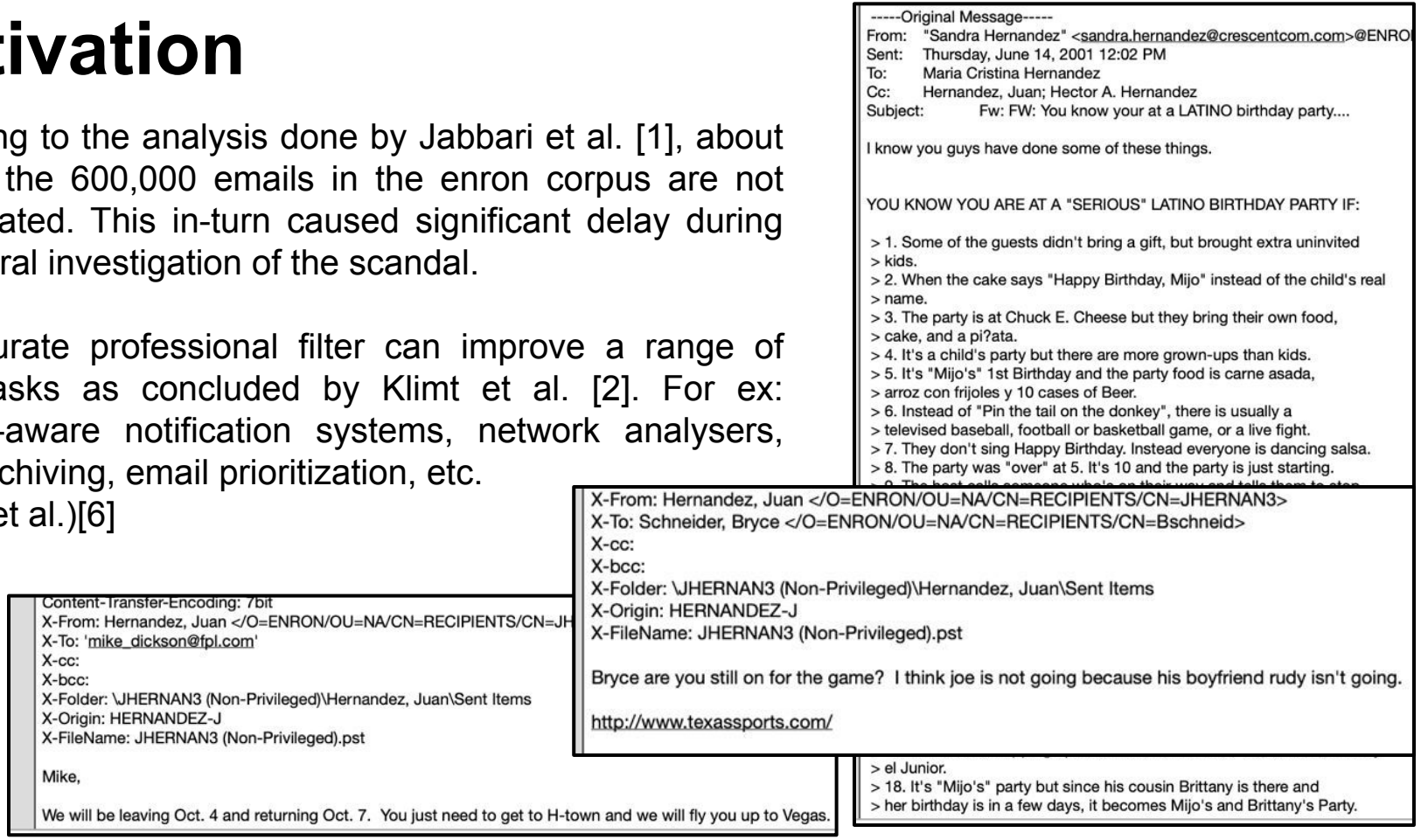
Introduction

Thanks to widespread use of emails, different research problems related to email classification have arisen. Ex. Spam filters, priority assignment, automated foldering, etc. For this project, we conducted a study on email classification into two categories: Work-related and Personal. We scrape a subset of the Enron emails corpus and engineer features to train a variety of ML models for the classification task.

Motivation

- According to the analysis done by Jabbari et al. [1], about 20% of the 600,000 emails in the enron corpus are not work-related. This in-turn caused significant delay during the federal investigation of the scandal.

- An accurate professional filter can improve a range of email tasks as concluded by Klimt et al. [2]. For ex: Context-aware notification systems, network analysers, email archiving, email prioritization, etc.
(Graus et al.)[6]



Overview

We manually annotated 5,028 samples of the enron email corpus for this task and extracted lexical and semantic features from the emails. Similar to the approach by Alkhereyf et al. [3], we trained a variety of models on the engineered dataset and evaluated the results. We also discussed the possible improvements and future prospects of this classification task.

Dataset: Enron Email Dataset (2015 version) https://www.cs.cmu.edu/~enron/

Sample size used: 5,028

Softwares/Tools used: Label Studio, VS Code, MS Office, Google Docs

Models trained: SVM, AdaBoost, Random Forest, ANN

GitHub repository: https://github.ncsu.edu/adongar/G6_ALDA_Project

Methodology

Data preparation: Training data was prepared in 3 phases

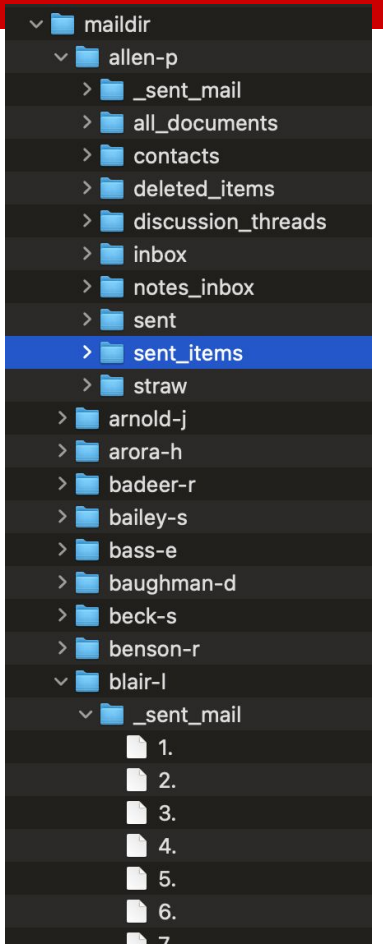
Phase 1: Crawling the folder structure according to the descriptions by Agarwal et al. [4] and Manual Labelling of samples into ‘work-related’ or not

Phase 2: Email Parsing and Feature Extraction

- Extract features like subject, domain name, no. of attachments, no. of recipients, length of the email, sent time, etc.
- Extract TF-IDF scores, and use GloVE 6B model to generate vector embeddings for the textual data.

message	label	sender_email	subject	num_receivers	email_length	email_domain	sent_time	is_forwarded

Extracted features

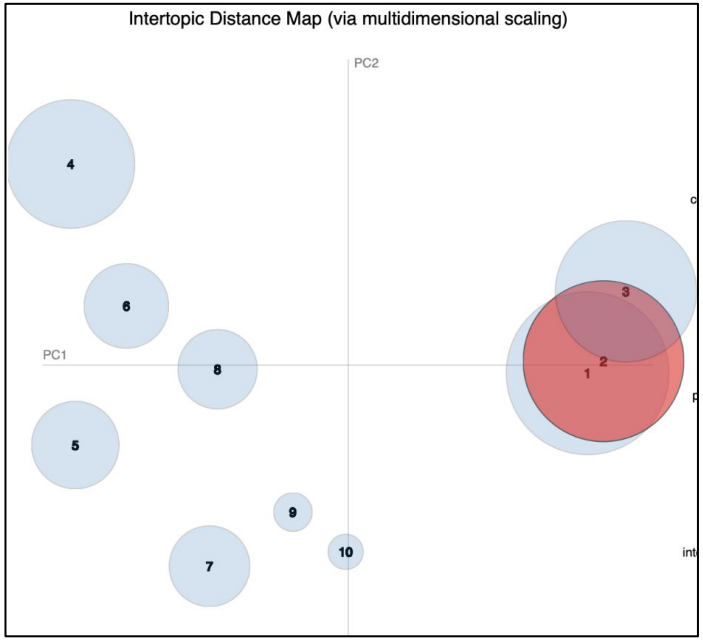


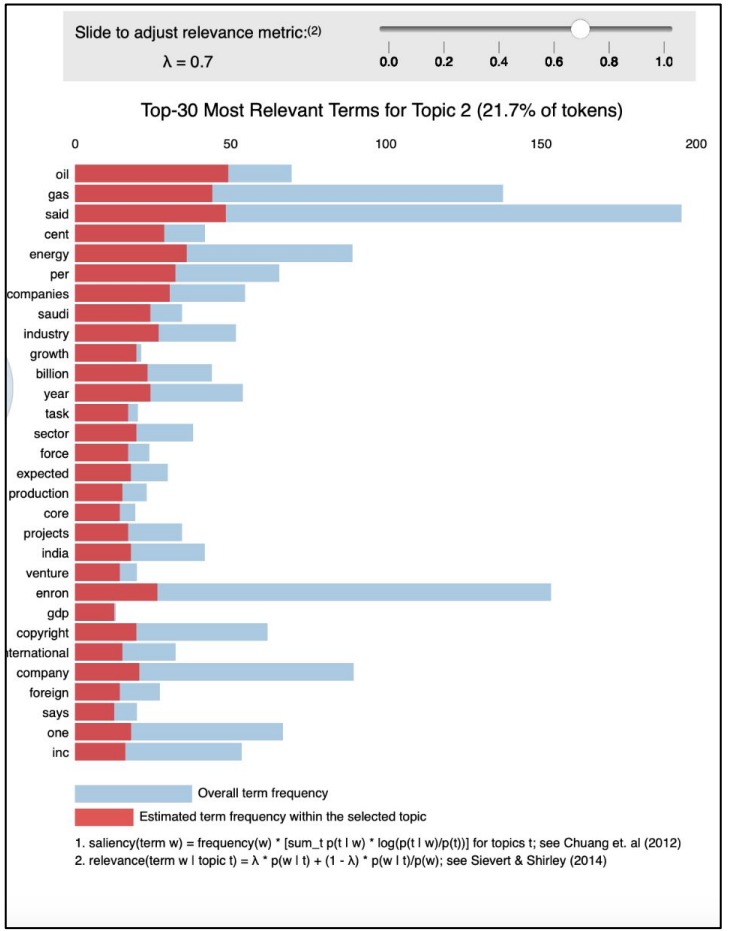
Enron Dataset Structure

Methodology

Phase 3: Topic Modelling and Data pre-processing

- Used Latent Dirichlet Allocation to assign dominant topics
- Condense the vector embeddings by frequency averaging
- Scaling and normalization of features





Methodology - ML Techniques

1. SVM

SVMs have shown excellent performance in email classification tasks. Also, email content typically results in high-dimensional feature spaces, where SVMs excel.

Kernel Type: Radial basis function

Regularization: 10

2. Random Forest

Random Forest can provide insights into which features are most important for classification, which is valuable for understanding work-related email characteristics.

n_estimators: 100

Goodness criteria: Gini impurity

Methodology - ML Techniques

3. AdaBoost

For spam classification, AdaBoost achieves 99.21% precision, outperforming other algorithms. It is equipped to distinguish between finer differences between work-related and non-work-related emails.

After Grid Search:

- n_estimators: 100
- Learning rate: 0.1

Cross-validation: 5-fold

4. Artificial Neural Networks

Neural networks can learn relevant features from raw email text, potentially capturing complex patterns. They are extremely versatile and have proven performance for classification tasks.

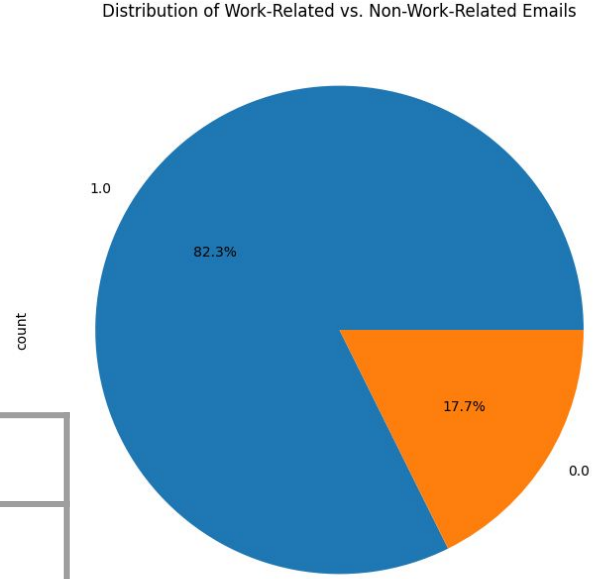
Input layer: Dense, 7 neurons, ReLU
Hidden Layer 1: Dense, 4 neurons, ReLU
Hidden Layer 2: Dense, 2 neurons, ReLU
Output Layer: Dense, Sigmoid
Optimizer: Adam, learning rate: 0.001
No. of epochs: 50 (with early stopping)

Results

Train-Test split: 80 : 20

Python version: 3.12.3

Majority class in training data: 82.3%



Model Name	Accuracy	Recall	F1 score
SVM	0.851	0.86	0.918
Random Forest	0.909	0.91	0.943
AdaBoost	0.849	0.85	0.791
ANN	0.848	0.86	0.82

Results - LLM

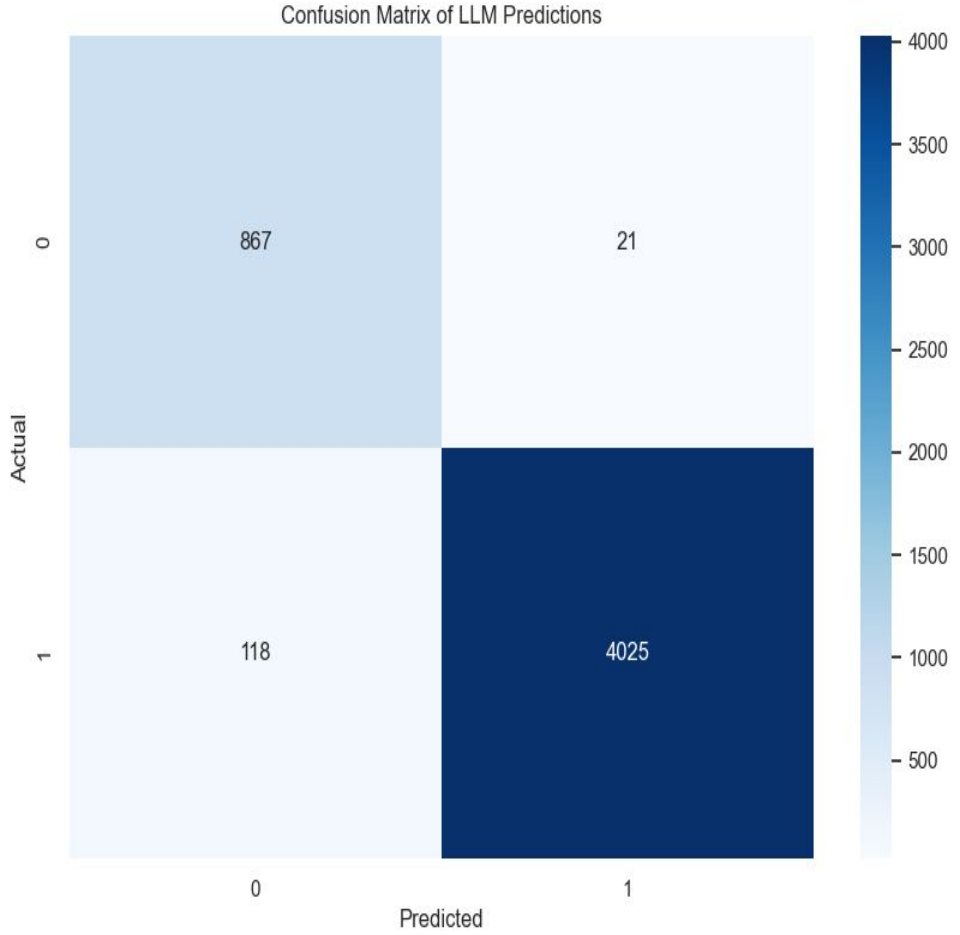
Model: Llama3-70b with Groq

- Hamming Loss Metric

Accuracy: 97.24%

F1 Score: 0.97

The prompt was framed based on the definitions of business related emails set by Alkhereyf et al. [3]



Discussion

1. **Random Forest:** This model performs the best across all metrics. This can be attributed to its ensemble nature, which combines multiple decision trees to make predictions.

2. **Support Vector Machine (SVM):** SVM shows the second-best performance, with high accuracy and recall (~0.86), and a particularly strong F1 score (0.918). This is likely due to its ability to find the optimal hyperplanes for separating different classes.

3. **AdaBoost & ANN:** These model shows comparable accuracy and recall to SVM but have notably lower F1 scores. They may be struggling to capture some of the nuances in the email data that Random Forest and SVM are able to identify. The lower F1 score for AdaBoost suggests it might be facing challenges in balancing false positives and false negatives.

Discussion

Earlier attempts to classify emails into ‘business’ & ‘personal’ classes (Gilbert et al. [5], Alkhereyf et al. [3]) try and model the social-network features of the enron dataset specifically. However, these are not global network features, and are very specific to the dataset samples. **Instead, by modelling the semantic features of the email, we can potentially achieve more generalizable results.**

			Business		
Trained on	Tested on	Accuracy	F-1	Recall	Precision
$\text{Enron}_{\cup\ tr}$	$\text{Enron}_{\cup\ ts}$	91.2	94.4	92.1	96.7
$\text{Enron}_{\cap A\ tr}$	$\text{Avocado}_{\cup\ ts}$	93.5	96.4	96.9	96.0

Conclusion

1. The results we achieved are considerable given the amount of labelled data. The models can be improved further by taking into consideration the class imbalance and more complex semantic features. In addition, we provided new annotated data samples as a supplementary contribution.

2. Furthermore, looking at the LLM's classifier's results, we can conclude that modern LLMs are extremely good at understanding context and perform very well at classification tasks similar to the one we undertook. Upon shifting the perspective, a new area of research is to evaluate the ability of LLMs to label datasets in a comparable way to humans. (Pavlovic et al.) [7]

References

1. 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 Lin- guistics, pages 407–411.

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

3. Work Hard, Play Hard: Email Classification on the Avocado and Enron Corpora" by Alkhereyf and Rambow. [http://www.cs.columbia.edu/nlp/papers/2017/alkhereyf_email_classification.pdf]. 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.

4. 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.

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

6. 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.

7. 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

8. "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora" by Bekkerman et al. [https://ciir.cs.umass.edu/pubfiles/ir-418.pdf]

THANK YOU!