Email classification using Python

Ayuka Mikami
University of Michigan
Industrial and Operations Engineering
ayukamik@umich.edu

***Abstract*—In this project, we built a classifier that takes an email as input and the label as output. By using this classifier, we can know the category and importance of an email before reading through it, making email management easier. The process is preprocessings the email text, extracting features, clustering the data, and constructing the model. As a result, we achieved 98% overall accuracy, with each cluster scoring over 97% on all evaluation metrics. However, there is an issue that the categories of each cluster are not clearly distinct. Improving the performance of data preprocessing, feature extraction, and clustering can be additional tasks.**

I. Introduction

This project aims to build a classifier that takes an email as input, and the label as output. As a university student, I receive many emails every day, but do not always have time to read them carefully that can lead to missing important information. If we could know the category and importance of each email before reading it, it would be much easier to manage them. Email classification is a topic that a lot of research is being done[1]. Each of them has different approaches, but most of them focus on spam filtering. In this project, we will not define the number of clusters initially. Instead, we determine the optimal number of clusters while clustering the data, which is expected to uncover the natural structure of the dataset.

II. Method

The dataset used is available on the Hugging Face[2] which includes 19,600 emails that are 5,500 to 7,500 characters long. The overview of the process of this project is as follows. A. Data preprocessing and feature extraction, B. Clustering, C. Building the classifier

A. Data preprocessing and feature extraction

For data preprocessing, we will clean the email and make it into the form which is effective for feature extraction. First, we delete ”from”, ”sent”, ”to”, ”cc”, ”subject”, and the email addresses. We also remove symbols, except for ”.”, ”,”, ”!”, and ”?”. After extracting unnecessary parts and making the email clean enough, we do feature extraction. For this step, we use sentence embedding with sentence transformers. This method can vectorize the text while considering the meaning of the whole sentence not just each word.

B. Clustering

For the clustering, we use hierarchy clustering because the dataset has no labels and the number of clusters is unknown. First, we draw a dendrogram using the data, and pick an appropriate number of clusters. Then, label each data by applying hierarchy clustering.

C. Building the classifier

For the classifier, we use a Deep Neural Network(DNN) with two hidden layers. Input is the vectorized email, and output is the cluster label. We split the vectorized data into the training set(80%) and the test set(20%) for use. The loss function for this model is the cross entropy loss function, which is one of the standard approaches for classification. Also, since our data happened to be imbalanced, we will introduce the class weight to the loss function to consider that imbalance. The optimizer will be mini-batch gradient decent, which helps us find the optimal value while balancing the efficiency and the stability. After training the model, we do testing and evaluation. The evaluation will be based on overall accuracy, and precision, recall, and F1 score for each cluster.

III. Result

A. Data preprocessing and feature extraction

After data preprocessing, emails will be as shown in Table 1. Then we used sentence embedding for feature extraction, and construct a matrix with a vector for each email.

Table 1. Preprocessed email

Before	Harry – I think that we need to talk with EWS Tariff risk to make sure that we are on the same page. Please advise. Jim —*Original Message*— From: Ogenyi, Gloria Sent: Monday, November 19, 2001 3:27 PM To: Anderson, Bob; Collins, Patricia Cc: Rathvon, Richard; Sparling, Jay; Keene, Patrick; Kingerski, Harry; Steffes, James D.; Ryall, Jean Subject: RE: Rock Tenn CoGen Facility Bob, Per Section 25.345 (c) of the Electric rules...
After	Harry I think that we need to talk with EWS Tariff risk to make sure that we are on the same page. Please advise. Jim Original Message Bob, Per Section 25.345 c of the Electric rules...

B. Clustering

The dendrogram is shown in Figure 1. From the result, we set the number of clusters to 4 and applied hierarchy clustering. Example emails from each cluster are shown in

Table 2. From the result, we can assume that cluster 0 is advertisement, cluster 1 is important notification, cluster 2 is business emails, and cluster 3 is the others.

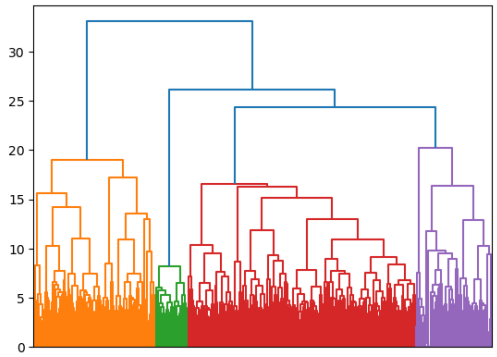

Fig. 1. Dendrogram

Table 2. Clustered emails

Cluster 0	**Email 1** The Valentines Day Countdown Has Begun Although you might have gotten the impression the holidays were over, the most important day of the year for lovers is just around the corner. At wine.com, we ... **Email 2** Make a Reservation On Time AWA is 1 America West ranks number one in ontime performance among all major carriers as recorded in the latest Air Travel Consumer Report published by the U.S. Department of Transportation...
Cluster 1	**Email 1** This is to update you on several CA bills that may have implications for green energyrenewables. Bills previously discussed SBX2 78 Status Last night the bill was approved by the State Assembly. Will be returned to Senate for reconciliation. Link httpinfo... **Email 2** By Jason Leopold LOS ANGELES Dow JonesFederal energy regulators may investigate whether the operator of Californias wholesale electricity market has been manipulating its powerpurchasing practices to reduce the costs of power...
Cluster 2	**Email 1** Hey Russell as well as Dan and Debra, Pursuant to our conversation of this morning, I just want to confirm the following. 1. We have agreed that we will use separate guarantees... **Email 2** Rick and David, Also give me your analysis wherein the Counterparty receivable financing is more generic, without specific references to contracts but a global reference to receivables...
Cluster 3	**Email 1** IMAGE June 12, 2001 Cant read this email? Click here Issue e10806 PROVANTAGE Customer To unsubscribe from the? **Email 2** As requested, your News Alert for YHOO follows from EquityAlert.com. To editdiscontinue your alerts please refer to end of message. PLEASE REVIEW THE NOTICE AND DISCLAIMER BELOW Paid Advertisement

C. Building the classifier

The performance of the classifier is shown in Table 3. As we can see, there is no variance between the metrics values for each cluster, and all data were above 97%. Some examples of misclassified emails are shown in Table 4. We can see emails with overlapping features among two clusters are likely to be misclassified.

Table 3. Model evaluation

Cluster	Precision	Recall	F1 Score	Accuracy
Cluster 0	0.9825	0.9912	0.9868	
Cluster 1	0.9961	1.0000	0.9980	0.9865
Cluster 2	0.9924	0.9840	0.9882	
Cluster 3	09713	0.9817	0.9765	

Table 4. Misclassified emails

Prediction	True lavel	Email text
3	2	COLUMBIA GAS TRANSMISSION CORPORATION NOTICE TO ALL INTERESTED PARTIES Effective Saturday, October 27, 2001 and Sunday, October 28, 2001, capacities will be as follows Excess MDWQ Available ISS Withdrawals Available SIT Withdrawals Available Imbalance Drawdowns Available
2	0	HoustonChronicle.com News Nov.19, 2001 Volume 6.47 In this Issue Letter from the Editor Plus Thanksgiving Day Parade Whats New at HoustonChronicle.com How to Send Us Community Notices How to Contact Us

IV. CONCLUSION

We were able to build a classifier with high performance, which takes an email as input and the label as output. This is achieving the goal of this project, although by seeing the email from each cluster, the category for each cluster was not obvious enough. Since the model training step achieved high performance, for future improvement, there are several possibilities in data preprocessing, feature extraction, and clustering. For data preprocessing, instead of removing the unnecessary part, extracting the main text part from the email may be valid. For feature extraction, dimensionality reduction techniques such as PCA might be effective in improving the efficiency of the data. For the clustering, we used hierarchy clustering since the number of clusters was unknown and we wanted to give some flexibility. By doing experiments for different numbers of clusters, trying other methods such as the Gaussian Mixture Model may help. By expanding this model and incorporating these improvements, we can expect to build a classification with higher performance and clear labeling.

REFERENCES

[1] Ghulam Mujtaba et al, “Email Classification Research Trends: Review and Open Issues”, IEEE Access, PP.1-1, 0.1109/ACCESS.2017.2702187.

[2] GalaktischeGurke/emails_5500_to_7500, HuggingFace, https://huggingface.co/datasets/GalaktischeGurke/emails_5500_to_7500