Project description: The aim is email spam classification. I have done ML techniques and want to try these methods.

**Part 1: Random string detection in spam email messages.**

The following can be done to achieve this via relative entropy…

* Compute entropy distribution of each of the dataset. Run classification task for each dataset.
* Generate equivalent number of random strings. And add to every dataset and then Measure entropy again.
* Re-run classification task adding one random string to each message (test set).
* Compare relative entropy for distributions before and after adding random string.

u can read this article for reference...

<https://redcanary.com/blog/threat-hunting-entropy/>

also this one... <https://medium.com/30-days-of-machine-learning/day-5-entropy-relative-entropy-and-cross-entropy-8369d67cc180>

Datasets-they are 4 datasets divided into training and testing and labelled as spam and ham emails, I will give them to you. so preprocessing on them has been done on these datasets.

OUTPUT- the output could be on the lines of below

I would need deliverables for each dataset ( 4 of them) as

a) entropy values and distributions details for each email and dataset, the classification results as accuracy, precision, f1 score etc

b) once you add the random strings to each email, the new entropy values and distributions, new classification results as fi score, accuracy, precision, true positives etc

c. the details stats on difference in two entropy distributions

d. all code files and help me with running them to duplicate the results myself

e. description of what you have done- process and methods along with which libraries used

**Part 2: email spam detection based on user profiling**

Firstly from the given datasets, we need to do user profiling to develop feature set based on the types of good (ham) emails in their inbox

**Part 2.1 User profiling -** if user does not normally get emails of a kind normally and now they are then these emails are possibly spam(unwanted) for them

To do this

This can be done by doing feature extraction from the training data of a particular user. Let me explain the theory behind this. All users receives emails... right?

but same email may be spam(unwanted) for one user and ham(wanted) for another based on their preference.

User profiling aims to understand the preferences of a user by applying ML to learn from the user inbox data. Then if a new email comes with a particular topic heading which is not at all matching with the features identified for that user it is potentially spam.

It's like some user likes war movies, some like romance, some like animation.

So if a user likes romance and animation movies and receives emails related to war movies then it is unwanted(spam) for that user.

Another benefit of user profiling is that most content based filters are generic and are applied same for all users. These filters remove some emails that are mass sent but may be ham(wanted) for this user or gets thrown into junk folder becos the mail server filter detected it as spam.

By profiling we can re-fileter the spam folder and bring those False positives( emails that are ham but false detected as spam) back into inbox

Dataset- For user profiling I have enron dataset in raw form where I have emails for particular users.

\*\*\*\*\*\*

Once profiling done then spam emails can be classified….

**Part 2.2- By topic identification/detector in emails**

By identifying a topic profiler for the users if a new email has no topic or unrelated topic then it could be spam or random, not important for user

For e.g. if user doesn't normally get viagra emails then reject!

To do this step:

I would want to take the topic or subject line of the email as the feature and develop function from training dateset based on the keywords found in the subject-line of the training data

we look for subject line and match with what has been learnt from the training data. If the new email subject-line or topic is different to that it is potential spam.

for e.g. if the training data for a user does not have much emails with blank subjectline or topic and new email comes with no topic then it is spam potentially

so we train model based on subject lines given in training mails and then use that model for new mails

Can be done using word embedding, reference below

<https://www.kaggle.com/nhrade/text-classification-using-word-embeddings>

<https://towardsdatascience.com/word-embeddings-for-sentence-classification-c8cb664c5029>