Spams vs Hams Emails Classifications

By Burhan Sultan Basit

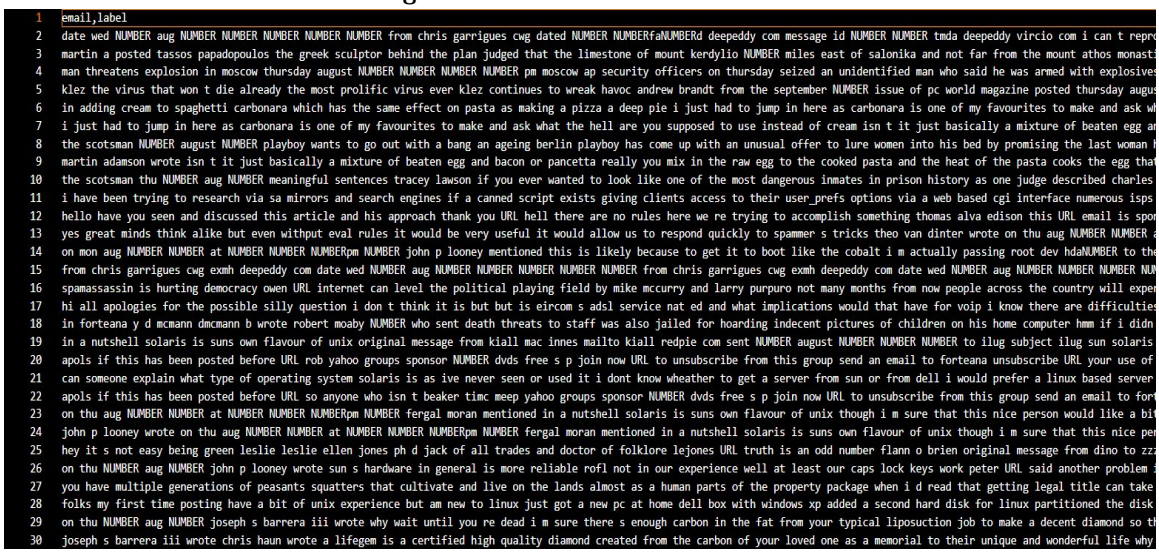
Student ID 650387980

1. Introduction

The goal of this project was to build a simple machine learning model that can detect whether an email is spam or not. This type of project is common because spam messages are a big problem in email systems. The work was done step by step, from cleaning the data, training the models, testing them, and finally creating a small app to make predictions.

2. Data Preparation

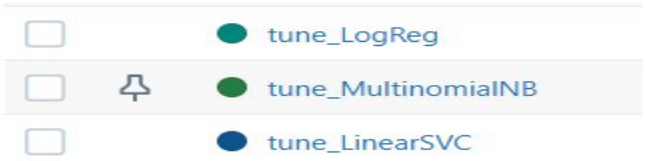
In this step, I worked with two different CSV files that contained email data. I checked for missing or repeated lines, and then joined them into one complete dataset. After combining them, I made sure the columns were clear , one called *email* and the other *label* (where 0 means ham and 1 means spam). I saved this final clean version as CLEAN_EMAILS.csv, which was later used for training the models.



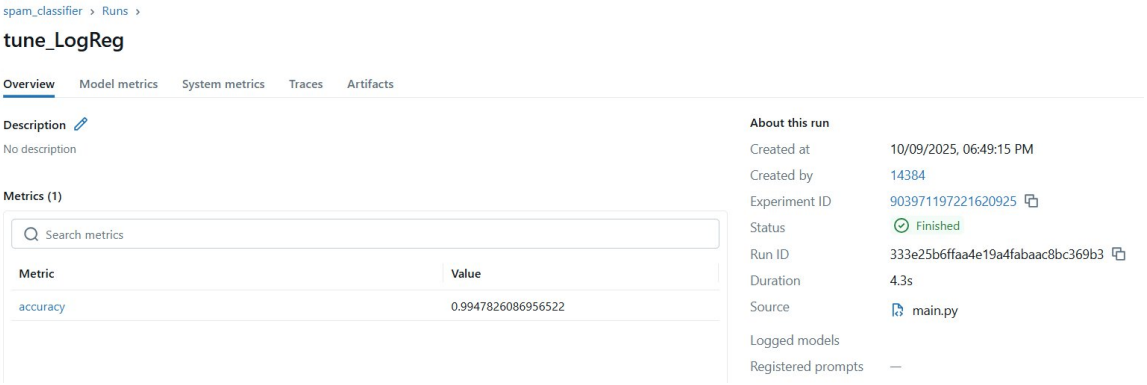
```
email,label
date wed NUMBER aug NUMBER NUMBER NUMBER NUMBER NUMBER from chris garrigues cwg dated NUMBER NUMBERfaNUMBERd deepeddy com message id NUMBER NUMBER tmda deepeddy vircio com i can t repro
martin a posted tassos papadopoulos the greek sculptor behind the plan judged that the limestone of mount kerdylio NUMBER miles east of salonika and not far from the mount athos monasti
man threatens explosion in moscow thursday august NUMBER NUMBER NUMBER NUMBER pm moscow ap security officers on thursday seized an unidentified man who said he was armed with explosives
klez the virus that won t die already the most prolific virus ever klez continues to wreak havoc andrew brandt from the september NUMBER issue of pc world magazine posted thursday augus
in adding cream to spaghetti carbonara which has the same effect on pasta as making a pizza a deep pie i just had to jump in here as carbonara is one of my favourites to make and ask wh
i just had to jump in here as carbonara is one of my favourites to make and ask what the hell are you supposed to use instead of cream isn t it just basically a mixture of beaten egg an
the scotsman NUMBER august NUMBER playboy wants to go out with a bang an ageing berlin playboy has come up with an unusual offer to lure women into his bed by promising the last woman h
martin adamson wrote isn t it just basically a mixture of beaten egg and bacon or pancetta really you mix in the raw egg to the cooked pasta and the heat of the pasta cooks the egg that
the scotsman thu NUMBER aug NUMBER meaningful sentences tracey lawson if you ever wanted to look like one of the most dangerous inmates in prison history as one judge described charles
i have been trying to research via sa mirrors and search engines if a canned script exists giving clients access to their user_prefs options via a web based cgi interface numerous isps
hello have you seen and discussed this article and his approach thank you URL hell there are no rules here we re trying to accomplish something thomas alva edison this URL email is spon
yes great minds think alike but even withput eval rules it would be very useful it would allow us to respond quickly to spammer s tricks theo van dinter wrote on thu aug NUMBER NUMBER a
on mon aug NUMBER NUMBER at NUMBER NUMBER NUMBERpm NUMBER john p looney mentioned this is likely because to get it to boot like the cobalt i m actually passing root dev hdaNUMBER to the
from chris garrigues cwg exmh deepeddy com date wed NUMBER aug NUMBER NUMBER NUMBER NUMBER NUMBER from chris garrigues cwg exmh deepeddy com date wed NUMBER aug NUMBER NUMBER NUMBER NUM
spamassassin is hurting democracy owen URL internet can level the political playing field by mike mccurry and larry purpuro not many months from now people across the country will exper
hi all apologies for the possible silly question i don t think it is but but is eircom s adsl service nat ed and what implications would that have for voip i know there are difficulties
in forteana y d mcmann dmcmann b wrote robert moaby NUMBER who sent death threats to staff was also jailed for hoarding indecent pictures of children on his home computer hmm if i didn
in a nutshell solaris is suns own flavour of unix original message from kiall mac innes mailto kiall redpie com sent NUMBER august NUMBER NUMBER NUMBER to ilug subject ilug sun solaris
apols if this has been posted before URL rob yahoo groups sponsor NUMBER dvds free s p join now URL to unsubscribe from this group send an email to forteana unsubscribe URL your use of
can someone explain what type of operating system solaris is as ive never seen or used it i dont know wheather to get a server from sun or from dell i would prefer a linux based server
apols if this has been posted before URL so anyone who isn t beaker timc meep yahoo groups sponsor NUMBER dvds free s p join now URL to unsubscribe from this group send an email to for
on thu aug NUMBER NUMBER at NUMBER NUMBER NUMBERpm NUMBER fergal moran mentioned in a nutshell solaris is suns own flavour of unix though i m sure that this nice person would like a bit
john p looney wrote on thu aug NUMBER NUMBER at NUMBER NUMBER NUMBERpm NUMBER fergal moran mentioned in a nutshell solaris is suns own flavour of unix though i m sure that this nice per
hey it s not easy being green leslie leslie ellen jones ph d jack of all trades and doctor of folklore lejones URL truth is an odd number flann o brien original message from dino to zzz
on thu NUMBER aug NUMBER john p looney wrote sun s hardware in general is more reliable rofl not in our experience well at least our caps lock keys work peter URL said another problem i
you have multiple generations of peasants squatters that cultivate and live on the lands almost as a human parts of the property package when i d read that getting legal title can take
folks my first time posting have a bit of unix experience but am new to linux just got a new pc at home dell box with windows xp added a second hard disk for linux partitioned the disk
on thu NUMBER aug NUMBER joseph s barrera iii wrote why wait until you re dead i m sure there s enough carbon in the fat from your typical liposuction job to make a decent diamond so th
joseph s barrera iii wrote chris haun wrote a lifegem is a certified high quality diamond created from the carbon of your loved one as a memorial to their unique and wonderful life why
```

3. Modeling and MLflow

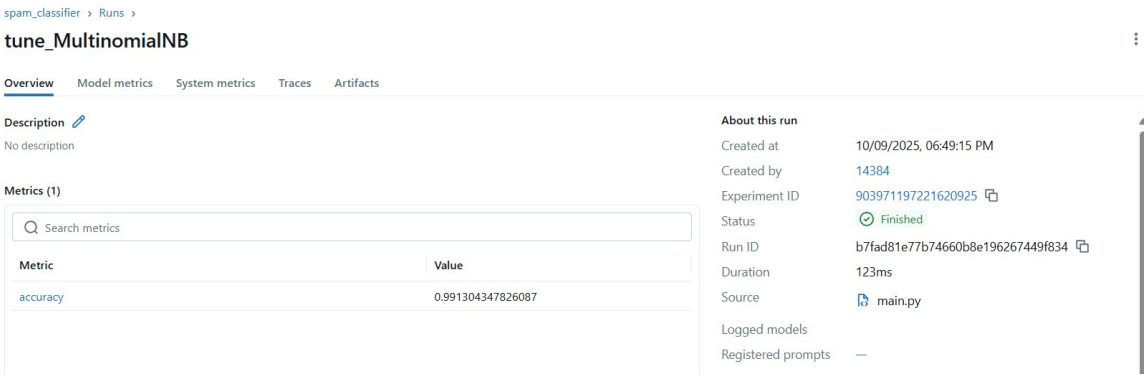
After preparing the clean data, I changed all the email text into numbers using a TF-IDF vectorizer so that the models could understand it. Then I trained three different models: Naive Bayes, Logistic Regression, and Linear SVM. Each model was tested and compared based on accuracy and F1 score. To keep everything organized, I used MLflow to track the results. It allowed me to see which model performed best and how much time each one took to train. Overall, the training went smoothly, and MLflow helped me visualize and compare all three runs in one place.



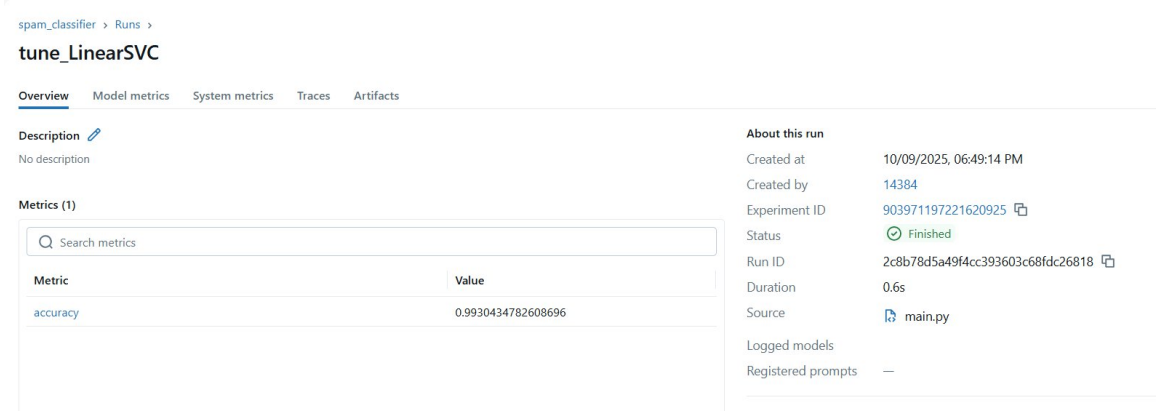
LogReg:



MultinimialNB:



Linearsvc:



4. Results

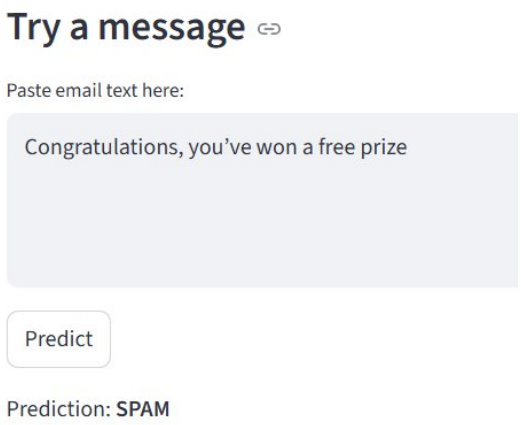
After training all three models, I compared their results. All of them gave good accuracy, but the Linear SVM and Logistic Regression models performed a little better than Naive Bayes. I checked the results using different metrics like Accuracy, Precision, Recall, and F1 Score. The confusion matrix also helped me see how many spam and ham emails were predicted correctly. Overall, the models were reliable and showed that the data cleaning and preprocessing were done properly.

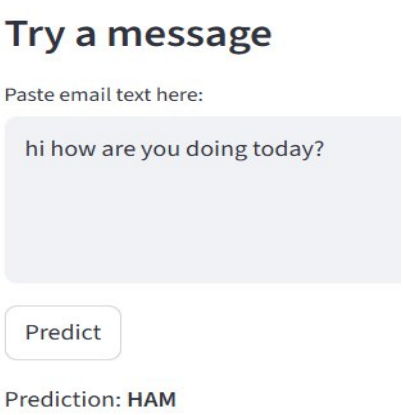
```
RESULTS (accuracy)
LinearSVC    0.9896
MultinomialNB 0.9930
LogReg       0.9913

BEST MODEL: MultinomialNB (acc=0.9930)

CLASSIFICATION REPORT (Ham=0, Spam=1):
              precision    recall  f1-score   support

         Ham     0.9939    0.9980    0.9959       489
        Spam     0.9881    0.9651    0.9765        86

    accuracy                         0.9930       575
   macro avg     0.9910    0.9815    0.9862       575
weighted avg     0.9930    0.9930    0.9930       575
```

5. Streamlit App

To make my project easy to use, I built a small Streamlit web app.
In the app, users can type or paste an email and see right away if it's spam or ham.
It also shows a confidence percentage, so you can know how sure the model is about its

prediction. The app also lets me choose between the three models (Naive Bayes, Logistic Regression, or Linear SVM) and compare how each one behaves.





6. Conclusion

This project helped me understand how machine learning can be used to detect spam emails.I did every step cleaning the data, building and testing models, comparing the results, and building an app. Using MLflow made it easier to track the experiments, and Streamlit made it possible to test the models interactively.
Overall, this project shows how data, code, and interface come together to solve a real problem.

Email Spam Classifier (Pick a Model)

Choose model:

- ○ LinearSVC
- ◉ MultinomialNB
- ○ LogisticRegression

Test accuracy: **0.9930**

Classification Report

	precision	recall	f1-score	support
Ham	0.9939	0.9980	0.9959	489
Spam	0.9881	0.9651	0.9765	86
accuracy			0.9930	575
macro avg	0.9910	0.9815	0.9862	575
weighted avg	0.9930	0.9930	0.9930	575

End of Report