Report For Machine Classification

1.Introduction

This a project focused on building a **machine learning classifier for text data**. The dataset, provided in the form of an email attachment, includes records with **no predefined labels**. The goal of the project is to identify and assign meaningful labels based on the content of the "Request Detail" , "Notes" fields, "Subject", fields, "Subject_Eng" fields and "Request Type" fields which are predominantly in Arabic, though English may also be present, in this project I used many advanced deep learning techniques and **labelling like Zero shot and Few shot classification** , in classification used many llm models like **BERT and variants** , and **Llama 3.2-instracut-1b** and **tries Gemma Bilingual.** and many Others…..

Code link: [githup link]

2.Preprocessing

I used the following techniques in addition to others for many texts:
Drop Unnamed: 0 column and Uneeded columns - Apply Stemming and LemmatizationRemove Null values - Remove Duplicate values - Apply Arabic normalization - Remove Special Characters - Remove Punctuations-Apply Regular Expression Techniques with Arabic LettersRemove Numbers - Apply splitting hashtag to words - Remove Arabic Stop Words - Clean hashtag-Remove English Stop Words - (Optional) Remove emoji - Remove Whitespace - Remove URLs-Remove HTML Tags - Remove Emails -Remove Phone Numbers-Remove Fax Numbers-(Optional) Remove Tweets - Remove Arabic Numbers - Remove Arabic Diacritics-(Optional) Remove Outliers-(Optional) Apply Translation if Columns after cleaning most Values ' ' or empty values

```
' نفيد سعادتكم بوجود مشكله جهاز عياده العظام الدخول وفتح الجهاز بيوزر موظف وزاره الصحه نوع الجهاز رقم الجهاز برجاء الاطلاع وتوجيه المختصين
لبدكم لحل المشكله ولكم جزيل الشكر مدير تقنيه المعلومات بمستشفي السليل العام منصور عبدالله الحابي تنبيه باخلاء المس،وليه الرساله ومرفقاتها معد
ه لاستخدام المرسل اليه المقصود بالرساله فقط تحتوي معلومات سريه محميه قانونيا تكن الشخص المقصود فانه يمنع منعا باتا عرض نشر استخدام مصرح لد
محتوي نرجو اخطار المرسل طريق الرد البريد الالكتروني واتلاف النسخ الموجوده لديك تعد التصريحات الاراء المذكوره الرساله بالمرسل تمثل وزاره الصح
ه تتحمل الوزاره مس،وليه الاضرار الناتجه فيروسات تحملها الرساله '

df["Request Detail"][0]

'*** This is an external email. Be Vigilant and take precautions.***_x000D_*** Do not click links or open attachments or reply unless you
recognize the sender and their email address, and you are expecting the email.***_x000D__x000D__x000D_نفيد سعادتكم بوجود مشكله في جهاز عي
ادة العظام_x000D_الصحة_x000D_وزاره موظف بيوزر وفتح الجهاز_x000D__x000D_نوع الجهاز_x000D_حيث انه لا يمكن الدخول : DEEL OPTIPLEX 3080_x000D_الجهاز_x000D_ رقم :
DLH0PJ3_x000D__x000D_المشكلة_x000D_لبدكم لحل وتوجيه المختصين_x000D_يرجاء الاطلاع ولكم جزيل الشكر_x000D__x000D_،،،_x000D__x000D__x000D__x000D_مدير تقنيه
ة_x000D_المعلومات بمستشفي السليل العام_x000D_منصور بن عبدالله الحابي_x000D_0556355578_x000D__x000D__x000D_[MANSOUR ABDULLAH AL-HABI (2)]_x000D__x0
00D__x000D__x000D__x000D__x000D__x000D_تنبيه بإخلاء المسئولية: هذه الرسالة ومرفقاتها معدة لاستخدام المُرسَل إليه المقصود بالرسالة فقط و قد تحتوي على
معلومات سرية أو محمية قانونيا. إن لم تكن الشخص المقصود، فإنه يُمنع منعا باتا أي عرض أو نشر أو استخدام غير مصرح به للمحتوى. نرجو إخطار المُر
سل عن طريق الرد على هذا ال...'
```

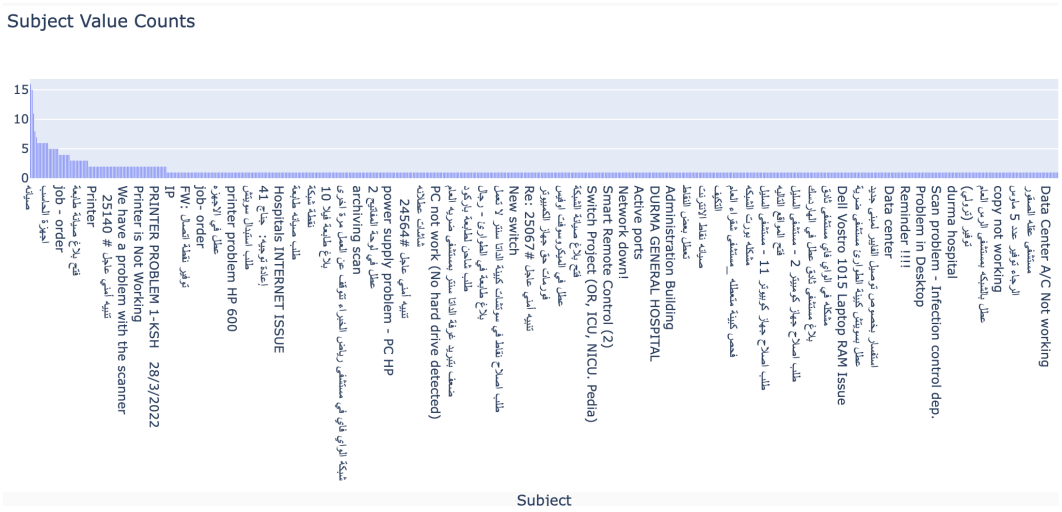
After & Before Processing for text

You can Check more about in the code in [githup]

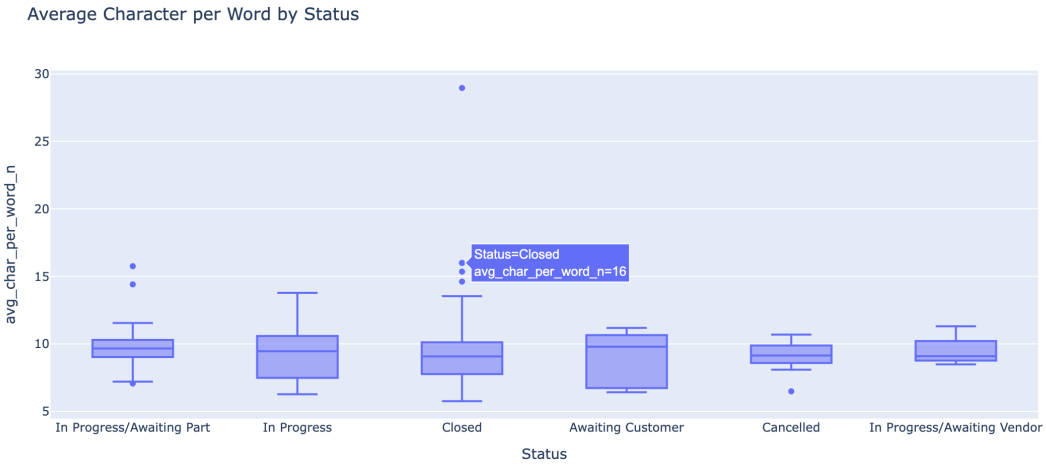
3.EDA

In EDA I used many libraries like matplotlib , plotly graph and express in addition to seaborn , for types of charts definitely I used them all ,bar chart, histogram, …… you can see the visualisation but here is a hint

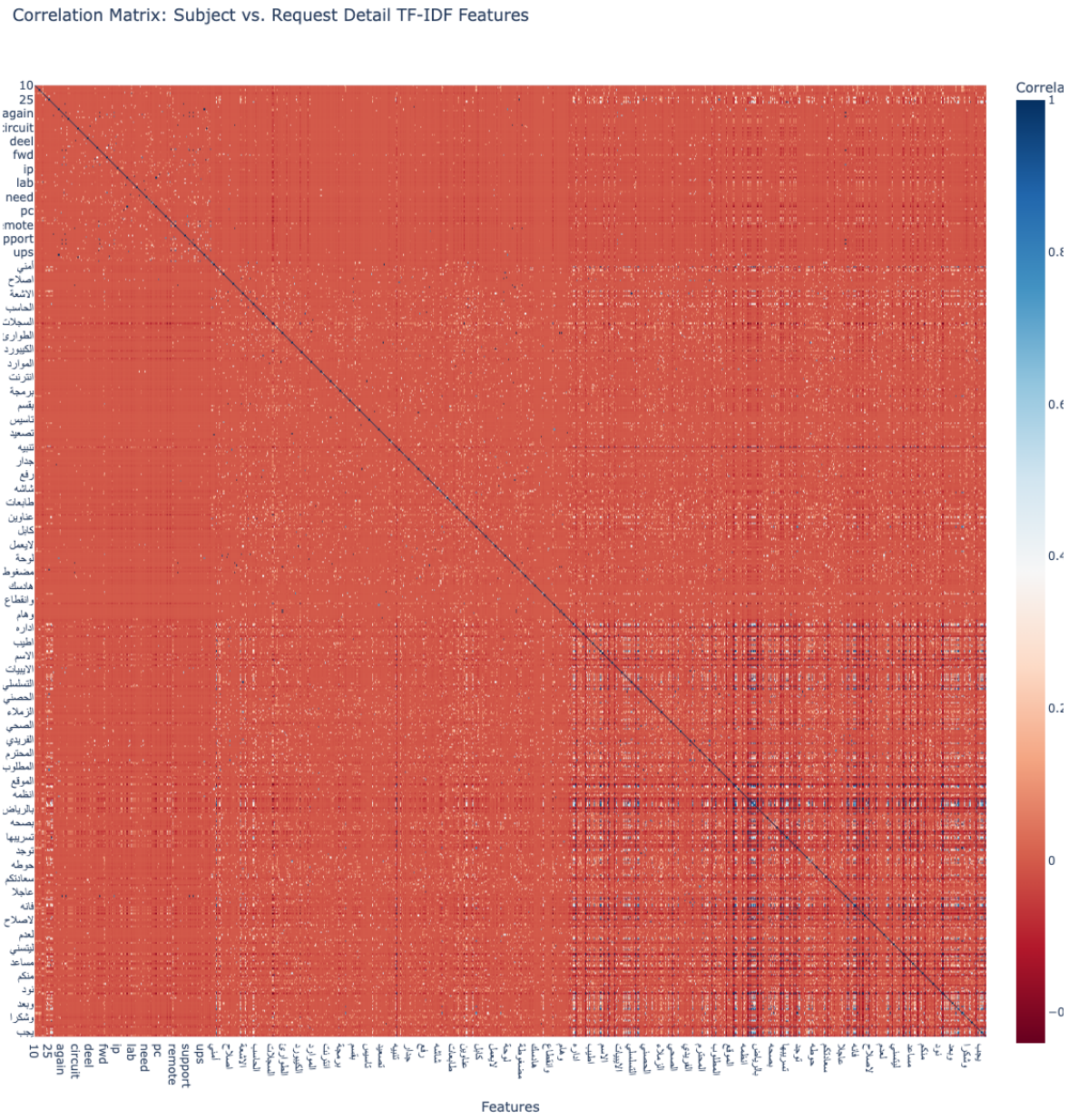
WordCloud For Combine Text of Request Detail and Notes



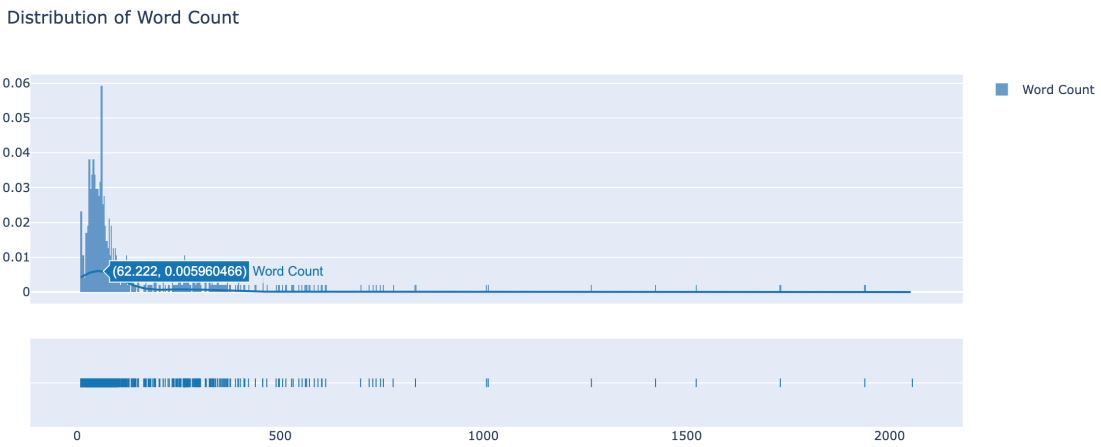
Counting for Subject in Our Data



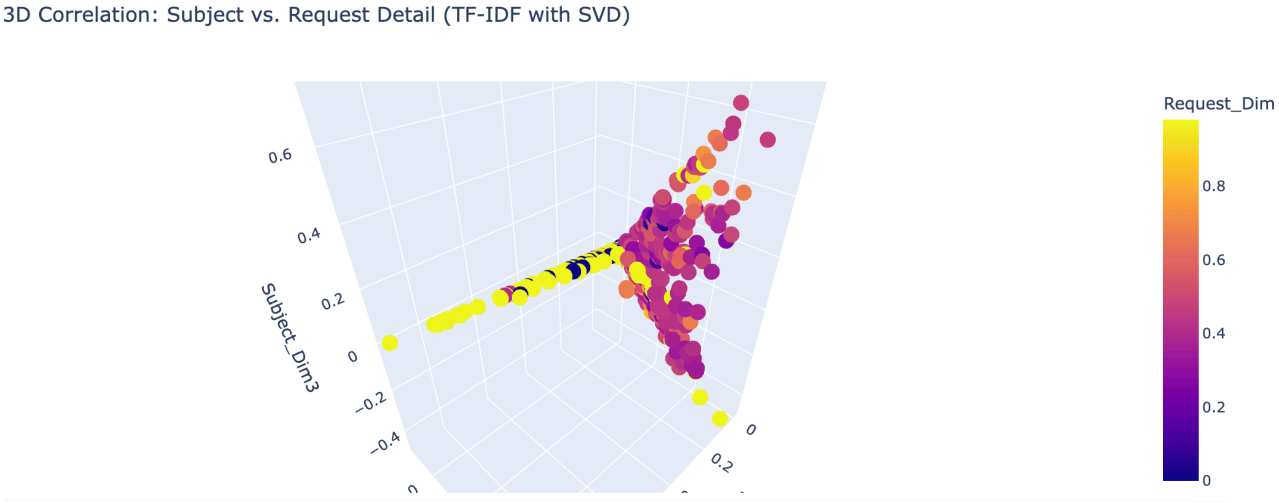
Box Plot for Average Character per Word declared by Staus



Correlation of Words between Subject And Request Detail



Distribution of Words with Interactive plotly Express



PCA with TF-IDF for Subject it's plot on 3D you can check it in the notebook

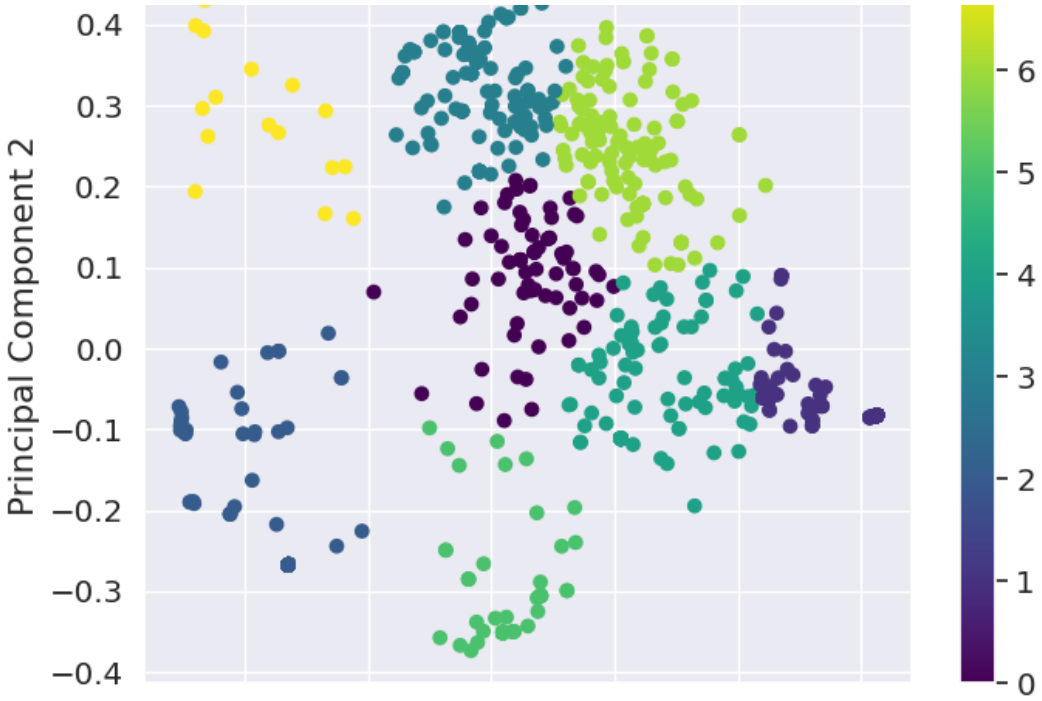
For more plots with apply the Explanation you can check in the notebook
Some

4.Labeling

- **Labelling with Translate and Upsample with FuzzWuzz**: Fuzzy string matching like a boss. It uses Levenshtein Distance to calculate the differences between sequences in a simple-to-use package. I have got the most frequent words. In Notes and Request Details and as results from that , I got **10 classes**

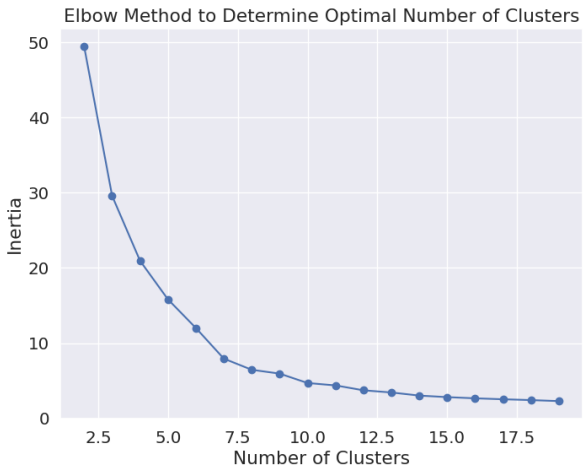
- **Apply Zero-shot classification with these xlm-roberta-large-xnli**: I have suggested about 6 classes from the my understanding to the data and git as as final result

- **Labelling using KMeans and PCA:** I have used TF-IDF with 100 features and and PCA with n_compoent:2 and use about 8 clusters and as result from it I got 8 classes



PCA & K-Means -> 8 Classes

- **Elbow Method to Determine Optimal Number of Clusters:** I have use these methods with kmean and range the cutters between 2:20 cluster and after that I choose the best number of clusters that do not make the Inertia expand as the relations ship between them is opposite and choose 5 cluster to 5 classes



Inertia with n-clusters

- **Zero-shot and Few Shot classification using T5**: As a candidate classes I required I choose about 5, 6 classes to base on them for the results , you can check the implementation in the code

5.Modeling

In 5 classes

1- model_name="asafaya/bert-mini-arabic"
Trained Text is A Combination fo Notes After transalte the needed infotmation from it , and Request Details Classes: label_list_HARD = ['Hardware Issue', 'Network Problem', 'Software Problem', 'Power Supply Issue', 'Peripheral Issue']

	precision	recall	f1-score	support
Hardware Issue	0.80	0.95	0.87	21
Network Problem	1.00	1.00	1.00	21
Peripheral Issue	1.00	1.00	1.00	7
Power Supply Issue	0.94	0.89	0.92	19
Software Problem	0.88	0.79	0.84	29
accuracy			0.91	97
macro avg	0.93	0.93	0.92	97
weighted avg	0.91	0.91	0.91	97

MiniAraBert

2- .model_name ="CAMeL-Lab/bert-base-arabic-camelbert-ca-poetry"
Trained Text is A Combination fo Notes After transalte the needed infotmation from it , and Request DetailsMapping for these Part of Classes using PCA for it wiht kmeans and give a range from 2:20 to get the best and get the best intria_ at kmean=5
Classes: label_list_HARD = ['Hardware Issue', 'Network Problem', 'Software Problem', 'Power Supply Issue', 'Peripheral Issue']

Training Loss	Validation Loss	Macro F1	Macro Precision	Macro Recall	Accuracy
No log	0.961537	0.533845	0.690942	0.536550	0.695876
No log	0.416179	0.718888	0.713822	0.729676	0.876289
No log	0.317929	0.866053	0.897036	0.861965	0.881443
No log	0.259281	0.899860	0.882874	0.926959	0.902062
No log	0.260063	0.909446	0.893069	0.935569	0.912371
No log	0.211095	0.900571	0.884024	0.926959	0.902062

```
utput(global_step=240, training_loss=0.3965223948160807, metrics={'train_runtime': 297.2326, 'train_samples_per_second': 2
_steps_per_second': 0.807, 'total_flos': 498082643684352.0, 'train_loss': 0.3965223948160807, 'epoch': 9.795918367346939})
```

Training Logs

I have try the training with KFold and get the following results

	precision	recall	f1-score	support
Hardware Issue	0.00	0.00	0.00	29
Network Problem	0.93	0.90	0.92	42
Peripheral Issue	0.00	0.00	0.00	11
Power Supply Issue	1.00	0.43	0.61	46
Software Problem	0.50	1.00	0.67	66
accuracy			0.64	194
macro avg	0.49	0.47	0.44	194
weighted avg	0.61	0.64	0.57	194

KFold-Classification Report

10 Classes

3.model_name="asafaya/bert-mini-arabic" -> Meduim
Just Use the Request Details as my main sentences
Use FuzzyWuzz Methods for making the classes and choose the nest part of it when number of classes = 10 in addition I balance the classes
Classes: label_list_HARD = [Printer Issues, Uncategorized, Miscellaneous, Computer Issues , Network Issues , Maintenance Requests , General Hardware Issues , Peripheral Device Issues , Security Alerts , Job Orders]

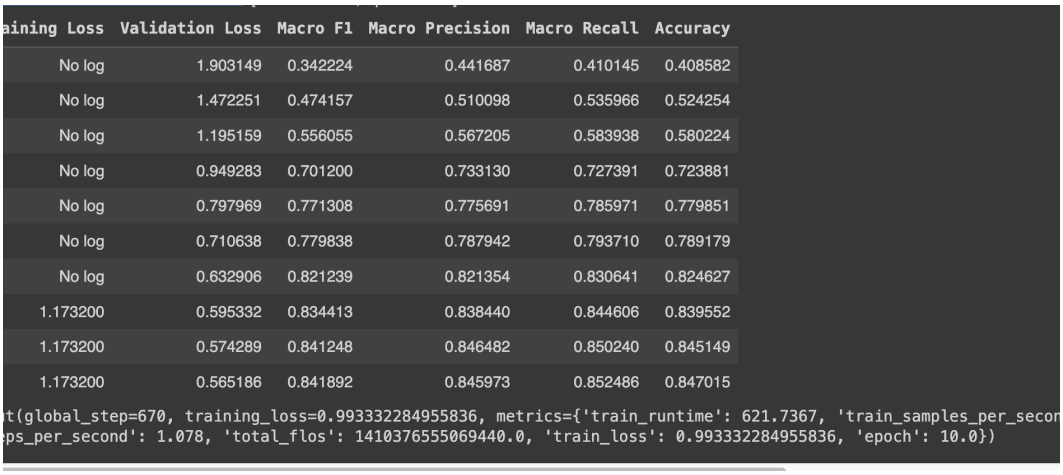
	precision	recall	f1-score	support
Computer Issues	0.92	0.75	0.83	32
General Hardware Issues	1.00	0.42	0.60	33
Job Orders	1.00	0.24	0.38	21
Maintenance Requests	0.22	0.92	0.35	24
Miscellaneous	0.84	0.70	0.76	23
Network Issues	1.00	0.54	0.70	26
Peripheral Device Issues	1.00	0.46	0.63	28
Printer Issues	0.43	0.48	0.45	25
Security Alerts	1.00	0.73	0.85	30
accuracy			0.59	242
macro avg	0.82	0.58	0.62	242
weighted avg	0.84	0.59	0.63	242

10 -Classes with Medium BERT

```
Average ROUGE-1: 0.5867768595041323
Average ROUGE-2: 0.0
Average ROUGE-L: 0.5867768595041323
Average BLEU Score: 1.0690088533109O6e-231
```

Rouge & BELU results

4.model_name ="CAMeL-Lab/bert-base-arabic-camelbert-ca-poetry"
Just Use the Request Details as my main sentences
Use FuzzyWuzz Methods for making the classes and choose the nest part of it when number of classes = 10 in addition I balance the classes
Classes: label_list_HARD = [Printer Issues, Uncategorized, Miscellaneous, Computer Issues , Network Issues , Maintenance Requests , General Hardware Issues , Peripheral Device Issues , Security Alerts , Job Orders]

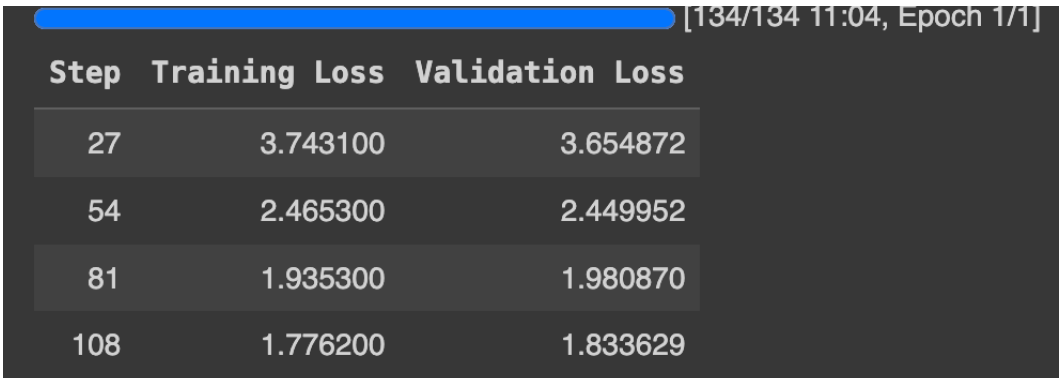


aining Loss	Validation Loss	Macro F1	Macro Precision	Macro Recall	Accuracy
No log	1.903149	0.342224	0.441687	0.410145	0.408582
No log	1.472251	0.474157	0.510098	0.535966	0.524254
No log	1.195159	0.556055	0.567205	0.583938	0.580224
No log	0.949283	0.701200	0.733130	0.727391	0.723881
No log	0.797969	0.771308	0.775691	0.785971	0.779851
No log	0.710638	0.779838	0.787942	0.793710	0.789179
No log	0.632906	0.821239	0.821354	0.830641	0.824627
1.173200	0.595332	0.834413	0.838440	0.844606	0.839552
1.173200	0.574289	0.841248	0.846482	0.850240	0.845149
1.173200	0.565186	0.841892	0.845973	0.852486	0.847015

t(global_step=670, training_loss=0.993332284955836, metrics={'train_runtime': 621.7367, 'train_samples_per_secon
ps_per_second': 1.078, 'total_flos': 1410376555069440.0, 'train_loss': 0.993332284955836, 'epoch': 10.0})

Training Loss for 10 Classes

5. Llama 3.2-instracut 1b for Text Classification
ust Use the Request Details as my main sentences
Use FuzzyWuzz Methods for making the classes and choose the nest part of it when number of classes = 10 in addition I balance the classes
Classes: label_list_HARD = [Printer Issues, Uncategorized, Miscellaneous, Computer Issues , Network Issues , Maintenance Requests , General Hardware Issues , Peripheral Device Issues , Security Alerts , Job Orders]



[134/134 11:04, Epoch 1/1]

Step	Training Loss	Validation Loss
27	3.743100	3.654872
54	2.465300	2.449952
81	1.935300	1.980870
108	1.776200	1.833629

Training Logs for One Epoch

```
Accuracy for label Computer Issues: 0.000
Accuracy for label Peripheral Device Issues: 0.000
Accuracy for label Security Alerts: 0.000
Accuracy for label Uncategorized: 0.848
Accuracy for label Printer Issues: 0.000
Accuracy for label Job Orders: 0.000
Accuracy for label Network Issues: 0.000
Accuracy for label Miscellaneous: 0.515
Accuracy for label Maintenance Requests: 0.000
Accuracy for label General Hardware Issues: 0.000

Classification Report:
                         precision    recall  f1-score   support

         Computer Issues       0.00      0.00      0.00        26
Peripheral Device Issues       0.00      0.00      0.00        26
         Security Alerts       0.00      0.00      0.00        24
           Uncategorized       0.16      0.85      0.27        33
          Printer Issues       0.00      0.00      0.00        20
              Job Orders       0.00      0.00      0.00        21
          Network Issues       0.00      0.00      0.00        31
           Miscellaneous       0.19      0.52      0.28        33
    Maintenance Requests       0.00      0.00      0.00        26
 General Hardware Issues       0.00      0.00      0.00        28

               micro avg       0.17      0.17      0.17       268
               macro avg       0.04      0.14      0.05       268
            weighted avg       0.04      0.17      0.07       268
```

Classification Report for the Trainings

```
Average ROUGE-1: 0.16417910447761194
Average ROUGE-2: 0.0
Average ROUGE-L: 0.16417910447761194
Average BLEU Score: 2.991067445358022e-232
```

Rouge & BELU for Llama3.2 -B Instruct for 1 Epoch

In addition for that I choose the **best from Every fine-tune model** based on that , and many model I wasn't have the computational cost for it like **gemma multilang** that I have try it for 2 days but nothing out I have but it also in the code in github

6.Evalution

For 5 classes

---

pipe("نفيد سعادتكم بوجود مشكله جهاز عياده العظام الدخول وفتح الجهاز بيوزر موظف نوع الجهاز رقم الجهاز برجاء الاطلاع وتوجيه المختصين ليدكم لحل المشكله ولكم جزيل الشكر مدير تقنيه المعلومات بمستشفي السليل العام منصور عبدالله الحابي تنبيه باخلاء المسءوليه ومرفقاتها معده لاستخدام اليه بالرساله فقط تحتوي سريه محميه قانونيا تكن الشخص فانه يمنع منعا باتا عرض نشر استخدام مصرح للمحتوي اخطار الالكتروني واتلاف النسخ الموجوده لديك تعد التصريحات الاراء بالمرسل تمثل تتحمل مسءوليه الاضرار الناتجه فيروسات تحملها")

Results: [[{'label': 'Hardware Issue', 'score': 0.0130720008164644 24},
{'label': 'Network Problem', 'score': 0.0092654563486576 08},
{'label': 'Software Problem', 'score': 0.0158361028879880 9},
{'label': 'Power Supply Issue', 'score': 0.9559705257415771},
{'label': 'Peripheral Issue', 'score': 0.0058559142053127 29}]]

For 10 Classes

pipe("نفيد سعادتكم بوجود مشكله جهاز عياده العظام الدخول وفتح الجهاز بيوزر موظف نوع الجهاز رقم الجهاز برجاء الاطلاع وتوجيه المختصين ليدكم لحل المشكله ولكم جزيل الشكر مدير تقنيه المعلومات بمستشفي السليل العام منصور عبدالله الحابي تنبيه باخلاء المسءوليه ومرفقاتها معده لاستخدام اليه بالرساله فقط تحتوي سريه محميه قانونيا تكن الشخص فانه يمنع منعا باتا عرض نشر استخدام مصرح للمحتوي اخطار الالكتروني واتلاف النسخ الموجوده لديك تعد التصريحات الاراء بالمرسل تمثل تتحمل مسءوليه الاضرار الناتجه فيروسات تحملها")

Results:
[[{'label': 'Computer Issues', 'score': 0.140791222453117 37},
{'label': 'Peripheral Device Issues', 'score': 0.004707982297986746},
{'label': 'Security Alerts', 'score': 0.019273553043603897},
{'label': 'Uncategorized', 'score': 0.05770004913210869},
{'label': 'Printer Issues', 'score': 0.05498286709189415},
{'label': 'Job Orders', 'score': 0.006223469041287899},
{'label': 'Network Issues', 'score': 0.014804959297180176},
{'label': 'Miscellaneous', 'score': 0.10969039797782898},
{'label': 'Maintenance Requests', 'score': 0.5517280697822571},
{'label': 'General Hardware Issues', 'score': 0.040097422897815704}]]

6.Conclusion

In this evaluation, I employed a lightweight model that demonstrated strong performance and is well-suited for deployment in a machine-based ticket support system. While more advanced models like GPT, LLaMA, or GEMMA variants (with over 7 billion parameters) offer higher capacity, they also come with significant computational overhead. For a task of this nature, which primarily involves text classification, such heavy models are not necessary.

The BERT-based models used in this pipeline achieved satisfactory accuracy and evaluation metrics, making them an ideal balance between performance and resource efficiency. Therefore, deploying these lightweight BERT variants provides a practical and cost-effective solution without sacrificing accuracy or reliability in the classification task