YOU TYPE, I TALK

Implementing a Natural Language Processing Voice Assistant

Group 8 – DREAM TEAM

Daniel Masamba & Kofi Nketia Ackaah-Gyasi

Link to github: https://github.com/kofinketia/NLP_project

Introduction

Background & Motivation

- The process of obtaining speech and text input, processing such input and giving information back either in the form of text or speech is a growing and fascinating aspect of Artificial Intelligence. As a group we have decided to focus on making meaning out of text input and responding with speech to make a conversation.
- Software application systems such as chatGPT, Siri, and Alexa, are great inspirations and resources which we are going to explore as we implement our project.
- We referenced the concepts of Natural Language Processing and Machine Learning, to design a web application capable of producing meaningful and appropriate speech information.

Problem Definition

- The area of speech and text processing in Artificial Intelligence is very fascinating
- However, how to go about such a project becomes daunting, as different models are needed as well as huge dataset for efficient learning
- What therefore do we seek to achieve?
- We want to make a simple web application where the user feeds an input text, and the model processes a speech response to the question provided.
- Our model should also be able to read and speak out the text provided
- A second major capability of our model is Language detection.
- Given a language in a particular context, the model identifies what the language is, and reads whatever foreign language is provided
- Pretty cool right☺

Challenges

- Collecting data was a major challenge
- Initial poor predictions by the model called for collection of more data to improve performance
- Our goal was to process both speech and text input. Processing speech input was very daunting and something we're still working on. However, speech output was completely successful.
- Not enough experience with NLP, we learned on the go
- Those models require a lot of resources that we did not have so we had to purchase Google Colab pro.

Data

Data Sources and Processing

The major aspect of this project was getting the data.

We collected common texts from 30 different languages, so that our model can learn the sequence of tokens in the language, and the respective language as well

The different languages include English, French, Arabic, Chinese, Korean, Dutch, German, etc.

We also collected conversational data mostly in English, and French, as well as translations between Languages

	Text	Language
13	胡赛尼本人和小说的主人公阿米尔一样，都是出生在阿富汗首都喀布尔，少年时代便离开了这个国家。胡...	Chinese
110	年月日，參與了「snh第三屆年度金曲大賞best」。月日，出演由优酷视频，盟将威影视，嗨乐...	Chinese
122	在他们出发之前，罗伯特·菲茨罗伊送给了达尔文一卷查尔斯·赖尔所著《地质学原理》（在南美他得到...	Chinese
151	系列的第一款作品《薩爾達傳說》（ゼルダの伝説）在年月日於日本發行，之後在年內於美國和歐洲地區...	Chinese
227	历史上的柔远驿是为了给琉球贡使及随员提供食宿之所，同时它也成为中琉间商业和文化交流的枢纽。琉...	Chinese
...	...	...
21880	这次计划的失败被归咎于军方高层之间缺乏沟通，其中所导致的最严重的问题就是让反动军在登陆时完全...	Chinese
21917	芝加哥大学主校区位于芝加哥市南的海德公园（hyde park）和伍德朗（woodlawn）街...	Chinese
21925	《追风筝的人》（英语：the kite runner，又译“追风筝的孩子”）是美籍阿富汗裔作...	Chinese
21982	戈尔巴乔夫结束了苏共的专制和暴政，使人民获得了民主、法治和自由，并使得东欧国家自主发展，结束...	Chinese
21998	年月，當時還只有歲的她在美國出道，以mai-k名義推出首張英文《baby i like》，由...	Chinese

1000 rows × 2 columns

	Text	Language
14	한국에서 성씨가 사용되기 시작한 정확한 시기는 알 수 없으나 한자漢字 등 중국 문물...	Korean
25	효모는 세포 수준의 생물학에서 모델 생물의 첫 번째 본보기로 간주해도 좋을 것이다 ...	Korean
67	세월호 참사 초기 구조 당시에 현장지휘관 osc·on scene-commander인...	Korean
83	^ hwang sw et al direct activation of capsaici...	Korean
99	자사 노선은 도부 스카이트리라인 오시아게 - 도부 도부쓰코엔 사이 도부 철도 이세사...	Korean
...	...	...
21862	그러나 년대로 접어들면서 이야기는 달라졌다 렌도이로 구단주의 취임과 함께 신흥강호로...	Korean
21921	나이트 엘프의 마지막 여왕이다 아즈샤라의 아버지는 아들이 없어서 아즈샤라 공주가 여...	Korean
21926	히타치 제작소의 철도 차량 제작 시스템 "a-train"을 채용하였으며 차량 제작에...	Korean
21944	이 운동을 회의적으로 생각하는 이유들이 몇몇 존재한다 현재 un 공용어는 많은 나...	Korean
21976	세기부터 세기까지 고전 포르투갈어의 두 번째 시기는 대항해 시대와 맞물려 아시아와 ...	Korean

1000 rows × 2 columns

	Text	Language
37	in johnson was awarded an american institute ...	English
40	bussy-saint-georges has built its identity on ...	English
76	minnesotas state parks are spread across the s...	English
90	nordahl road is a station served by north coun...	English
97	a talk by takis fotopoulos about the internati...	English
...	...	...
21829	on march empty mirrors press published epste...	English
21879	he [musk] wants to go to mars to back up human...	English
21896	overall the male is black above and white belo...	English
21897	tim reynolds born december in wiesbaden germ...	English
21951	the total high school population was now appro...	English

	Text	Language
46	قبل عام بالضبط وبتاريخ أعلن البغدادي خطة هدم ...	Arabic
53	وعندما وصل جنود الحملة الفرنسية غرب مدينة الإس...	Arabic
54	كان بصوته الشجي المميز خطيباً مميزاً وصاحب مدر...	Arabic
71	ليلة مايو دخل مسلحون داعش مدينة الرمادي عاصم...	Arabic
113	كل درجة من خط الطول شبه مقسمة إلى دقيقة وينقس...	Arabic
...	...	...
21847	تالتفراوت هو دُوَّار يقع بجماعة تاديغوست إقليم...	Arabic
21882	كان ثمة نتوء واضحٌ في حدقتيه فلقب بالحدقي ولكن...	Arabic
21913	تميز تجربته الفنية بالتنوع والغزارة الفنية وال...	Arabic
21939	لم يكن مراد معتاداً على هذا النوع من المعيشة ب...	Arabic
21974	في محاولة لتقليل احتمالات عدم السيطرة على الحر...	Arabic

1000 rows × 2 columns

	Text	Language
12	association de recherche et de sauvegarde de l...	French
22	la chirurgie comprenant principalement lablati...	French
23	dès les années les communes voisines darnouvi...	French
26	au er avril les services asama sont actuellem...	French
30	lalimentation industrielle convient parfaiteme...	French
...	...	...
21885	le pays de france a connu une occupation humai...	French
21980	résistant et gaulliste il est garde des sceaux...	French
21981	notices dans des bases relatives au sport ass...	French
21984	le village est une station familiale de sports...	French
21995	hors du terrain les années et sont des année...	French

1000 rows × 2 columns

	Text	Language
21	en navidad de poco después de que interpretó ...	Spanish
115	según el censo de [] había personas residien...	Spanish
162	en la copa mundial de fútbol sub- de pitó los...	Spanish
191	ally y buttons encuentran el descodificador y ...	Spanish
195	los primeros habitantes se establecieron cerca...	Spanish
...	...	...
21934	el de octubre de turcios lima falleció carbo...	Spanish
21959	para colmo las tropas albanesas atacaban conti...	Spanish
21975	fue fundado el de octubre de el día de ese ...	Spanish
21983	el investigador ha recibido varios reconocimie...	Spanish
21997	con motivo de la celebración del septuagésimoq...	Spanish

1000 rows × 2 columns

	Text	Language
	de spons behoort tot het geslacht haliclona en...	Dutch
	e prinses was als erfgename van polen een goe...	Dutch
	eliège werd gemeenteraadslid en burgemeester...	Dutch
	tussen een kruising in het westen van bischofs...	Dutch
	washed my hands in muddy water is een countr...	Dutch
...	...	...
21895	petersen brak nu door in hollywood voor twenti...	Dutch
21960	larressore is een gemeente in het franse depar...	Dutch
21963	phocas fokkens was vanaf lid en mede-oprichte...	Dutch
21964	op dit moment staat de zombie walk van oktobe...	Dutch
21971	het gewone volk leeft in vrede en nog altijd w...	Dutch

1000 rows × 2 columns

	Text	Language
0	klement gottwaldi surnukeha palsameeriti ning ...	Estonian
1	sebes joseph pereira thomas på eng the jesuit...	Swedish
2	ถนนเจริญกรุง อักษรโรมัน thanon charoen krung ...	Thai
3	விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...	Tamil
4	de spons behoort tot het geslacht haliclona en...	Dutch
...	...	...
10332	ನಿಮ್ಮ ತಪ್ಪು ಏನು ಬಂದಿದೆಯೆಂದರೆ ಆ ದಿನದಿಂದ ನಿಮಗೆ ಒ...	Kannada
10333	ನಾರ್ಸಿಸಾ ತಾನು ಮೊದಲಿಗೆ ಹೆಣಗಾಡುತ್ತಿದ್ದ ಮಾರ್ಗಗಳನ್...	Kannada
10334	ಹೇಗೆ ' ನಾರ್ಸಿಸಿಸಮ್ ಈಗ ಮರಿಯನ್ ಅವರಿಗೆ ಸಂಭವಿಸಿದ ಎ...	Kannada
10335	ಅವಳು ಈಗ ಹೆಚ್ಚು ಚಿನ್ನದ ಬೈಡ್ ಬಯಸುವುದಿಲ್ಲ ಎಂದು ...	Kannada
10336	ಟೆರ್ರಿ ನೀವು ನಿಜವಾಗಿಯೂ ಆ ದೇವದೂತನಂತೆ ಸ್ವಲ್ಪ ಕಾಣು...	Kannada

32337 rows × 2 columns

	Situation	emotion	labels
0	I remember going to the fireworks with my best...	sentimental	Was this a friend you were in love with, or ju...
1	I remember going to the fireworks with my best...	sentimental	Where has she gone?
2	I remember going to the fireworks with my best...	sentimental	Oh was this something that happened because of...
3	I remember going to the fireworks with my best...	sentimental	This was a best friend. I miss her.
4	I remember going to the fireworks with my best...	sentimental	We no longer talk.
...	...	...	...
64631	I found some pictures of my grandma in the att...	sentimental	Yeah I found some old pictures of when us kids...
64632	I found some pictures of my grandma in the att...	sentimental	Yeah reminds me of the good old days. I miss ...
64633	I woke up this morning to my wife telling me s...	surprised	Oh hey that's awesome! That is awesome right?
64634	I woke up this morning to my wife telling me s...	surprised	That is awesome!!!! Congratulations!
64635	I woke up this morning to my wife telling me s...	surprised	It is soooo awesome. We have been wanting a b...

64636 rows × 3 columns

Name
Conversation.csv
dataset.csv
emotion-emotion_69k.csv
eng_-french.csv
Language Detection.csv
languages.csv
List of languages by total number of speake...
Results.csv

	English words/sentences	French words/sentences
0	Hi.	Salut!
1	Run!	Cours !
2	Run!	Courez !
3	Who?	Qui ?
4	Wow!	Ça alors !
...	...	...
175616	Top-down economics never works, said Obama. "T...	« L'économie en partant du haut vers le bas, ç...
175617	A carbon footprint is the amount of carbon dio...	Une empreinte carbone est la somme de pollutio...
175618	Death is something that we're often discourage...	La mort est une chose qu'on nous décourage sou...
175619	Since there are usually multiple websites on a...	Puisqu'il y a de multiples sites web sur chaqu...
175620	If someone who doesn't know your background sa...	Si quelqu'un qui ne connaît pas vos antécédent...

175621 rows × 2 columns

	question	answer
0	hi, how are you doing?	i'm fine. how about yourself?
1	i'm fine. how about yourself?	i'm pretty good. thanks for asking.
2	i'm pretty good. thanks for asking.	no problem. so how have you been?
3	no problem. so how have you been?	i've been great. what about you?
4	i've been great. what about you?	i've been good. i'm in school right now.
...	...	...
3720	that's a good question. maybe it's not old age.	are you right-handed?
3721	are you right-handed?	yes. all my life.
3722	yes. all my life.	you're wearing out your right hand. stop using...
3723	you're wearing out your right hand. stop using...	but i do all my writing with my right hand.
3724	but i do all my writing with my right hand.	start typing instead. that way your left hand ...

3725 rows × 2 columns

Methodology

Procedure

Two main models alongside different frameworks were used to accomplish these tasks

Identify the language given a text input and output the detected language as a speech

DistilBert model for sequence classification, specifically DistilBertForSequenceClassification, using Pytorch

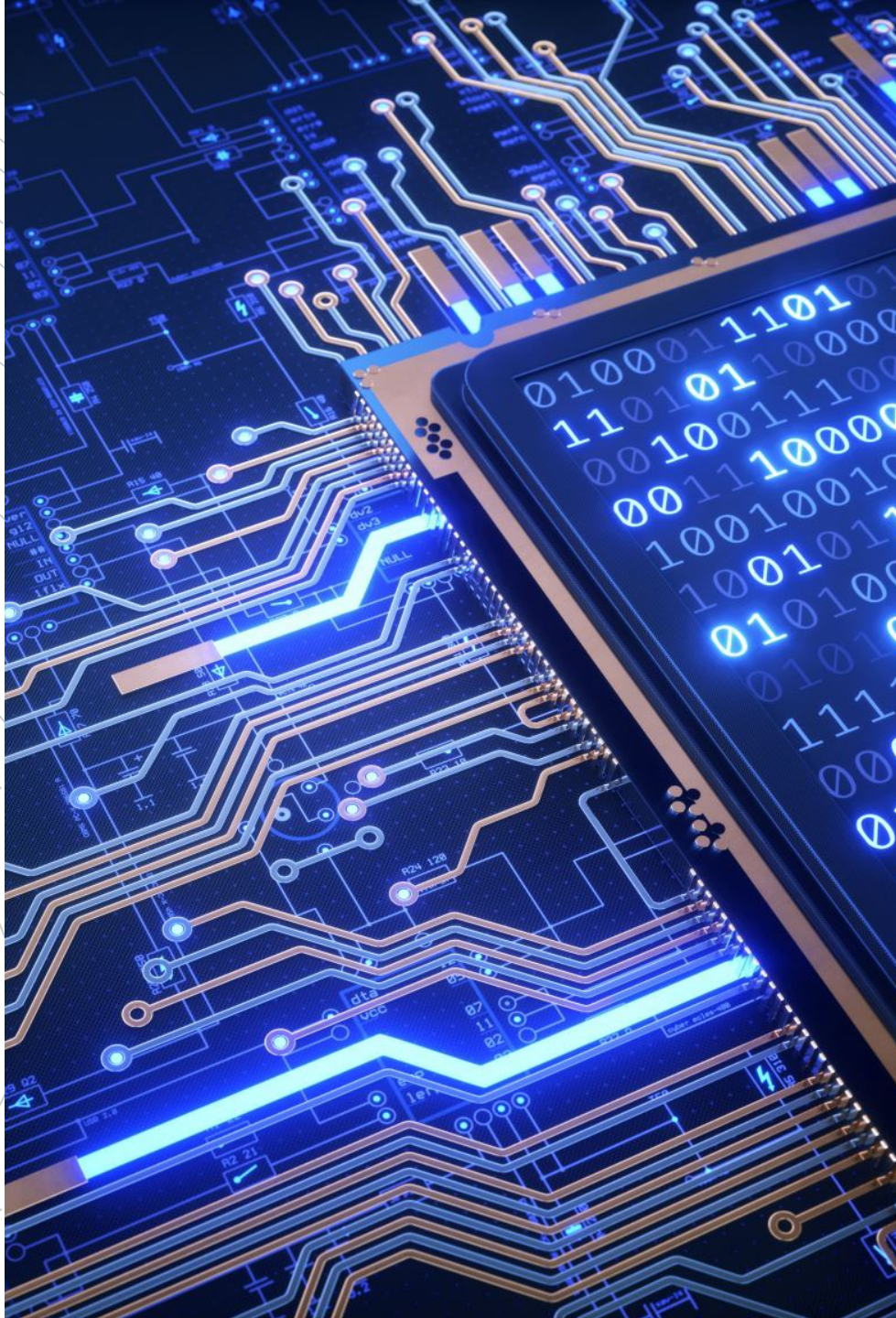
Added chatbot capabilities using the Generative Pre-trained Transformer (GPT) model from OpenAI, specifically GPTSimpleVectorIndex, that creates an index of documents used to perform similarity searches between documents.

sklearn's Label_Encoder was also integrated to encode the labels of text into a matrix

Google Text to Speech (gTTS), a google framework that supports converting text to speech and speech to text

Gradio, a web-based interface for testing machine learning models to interact with our model

Implementation



```
[ ]  # Convert language labels to integers
     label_encoder = LabelEncoder()
     df['label'] = label_encoder.fit_transform(df['Language'])

     # Split the dataset into train and validation sets
     train_texts, val_texts, train_labels, val_labels = train_test_split(
         df['Text'].to_numpy(),
         df['label'].to_numpy(),
         test_size=0.2,
         random_state=42)
```

```
[ ]  # Load the DistilBert tokenizer
     tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-multilingual-cased')

     # Tokenize the texts
     train_encodings = tokenizer(train_texts.tolist(), truncation=True, padding=True)
     val_encodings = tokenizer(val_texts.tolist(), truncation=True, padding=True)
```

	Text	Language	label
0	klement gottwaldi surnukeha palsameeriti ning ...	Estonian	5
1	sebes joseph pereira thomas på eng the jesuit...	Swedish	24
2	ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เ...	Thai	27
3	விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...	Tamil	26
4	de spons behoort tot het geslacht haliclona en...	Dutch	3
...	...	...	...
10332	ನಿಮ್ಮ ತಪ್ಪು ಏನು ಬಂದಿದೆಯೆಂದರೆ ಆ ದಿನದಿಂದ ನಿಮಗೆ ಒ...	Kannada	13
10333	ನಾರ್ಸಿಸಾ ತಾನು ಮೊದಲಿಗೆ ಹೆಣಗಾಡುತ್ತಿದ್ದ ಮಾರ್ಗಗಳನ್...	Kannada	13
10334	ಹೇಗೆ ' ನಾರ್ಸಿಸಿಸಮ್ ಈಗ ಮರಿಯನ್ ಅವರಿಗೆ ಸಂಭವಿಸಿದ ಎ...	Kannada	13
10335	ಅವಳು ಈಗ ಹೆಚ್ಚು ಚಿನ್ನದ ಬೈಡ್ ಬಯಸುವುದಿಲ್ಲ ಎಂದು ...	Kannada	13
10336	ಟೆರ್ರಿ ನೀವು ನಿಜವಾಗಿಯೂ ಆ ದೇವದೂತನಂತೆ ಸ್ವಲ್ಪ ಕಾಣು...	Kannada	13

32337 rows × 3 columns

```
[ ]  # Create the DataLoader for the training set
     train_dataset = TextDataset(train_encodings, train_labels)
     train_loader = DataLoader(train_dataset, batch_size=32, shuffle=True, collate_fn=DataCollatorWithPadding(tokenizer))

     # Create the DataLoader for the validation set
     val_dataset = TextDataset(val_encodings, val_labels)
     val_loader = DataLoader(val_dataset, batch_size=32, shuffle=False, collate_fn=DataCollatorWithPadding(tokenizer))
```

```
[ ]  # Load the DistilBertForSequenceClassification model
     model = DistilBertForSequenceClassification.from_pretrained('distilbert-base-multilingual-cased', num_labels=len(label_encoder.classes_))

     # Define the optimizer and the learning rate scheduler
     optimizer = torch.optim.AdamW(model.parameters(), lr=5e-5)
     total_steps = len(train_loader) * 3
     scheduler = get_linear_schedule_with_warmup(optimizer, num_warmup_steps=0, num_training_steps=total_steps)

     # Model summary
     device = torch.device('cuda') if torch.cuda.is_available() else torch.device('cpu')
     model.to(device)
```

```
DistilBertForSequenceClassification(
  (distilbert): DistilBertModel(
    (embeddings): Embeddings(
      (word_embeddings): Embedding(119547, 768, padding_idx=0)
      (position_embeddings): Embedding(512, 768)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (transformer): Transformer(
      (layer): ModuleList(
        (0-5): 6 x TransformerBlock(
          (attention): MultiHeadSelfAttention(
            (dropout): Dropout(p=0.1, inplace=False)
            (q_lin): Linear(in_features=768, out_features=768, bias=True)
            (k_lin): Linear(in_features=768, out_features=768, bias=True)
            (v_lin): Linear(in_features=768, out_features=768, bias=True)
            (out_lin): Linear(in_features=768, out_features=768, bias=True)
          )
          (sa_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (ffn): FFN(
            (dropout): Dropout(p=0.1, inplace=False)
            (lin1): Linear(in_features=768, out_features=3072, bias=True)
            (lin2): Linear(in_features=3072, out_features=768, bias=True)
            (activation): GELUActivation()
          )
          (output_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
        )
      )
    )
  )
  (pre_classifier): Linear(in_features=768, out_features=768, bias=True)
  (classifier): Linear(in_features=768, out_features=30, bias=True)
  (dropout): Dropout(p=0.2, inplace=False)
```

```python
# Train the model
train_loss_history = []
train_acc_history = []
val_loss_history = []
val_acc_history = []

for epoch in range(50):
    model.train()
    train_loss = 0
    total_correct = 0
    total_count = 0
    for batch in train_loader:
        batch = {k: v.to(device) for k, v in batch.items()}
        optimizer.zero_grad()
        outputs = model(**batch)
        loss = outputs.loss
        train_loss += loss.item() * batch['input_ids'].shape[0]
        loss.backward()
        optimizer.step()
        scheduler.step()
        _, predicted = torch.max(outputs.logits, 1)
        total_correct += (predicted == batch['labels']).sum().item()
        total_count += len(batch['labels'])

    train_accuracy = total_correct / total_count
    train_loss /= len(train_dataset)
    train_loss_history.append(train_loss)
    train_acc_history.append(train_accuracy)
    print(f"Epoch {epoch} - Train loss: {train_loss:.3f}, Train accuracy: {train_accuracy:.3f}")
```

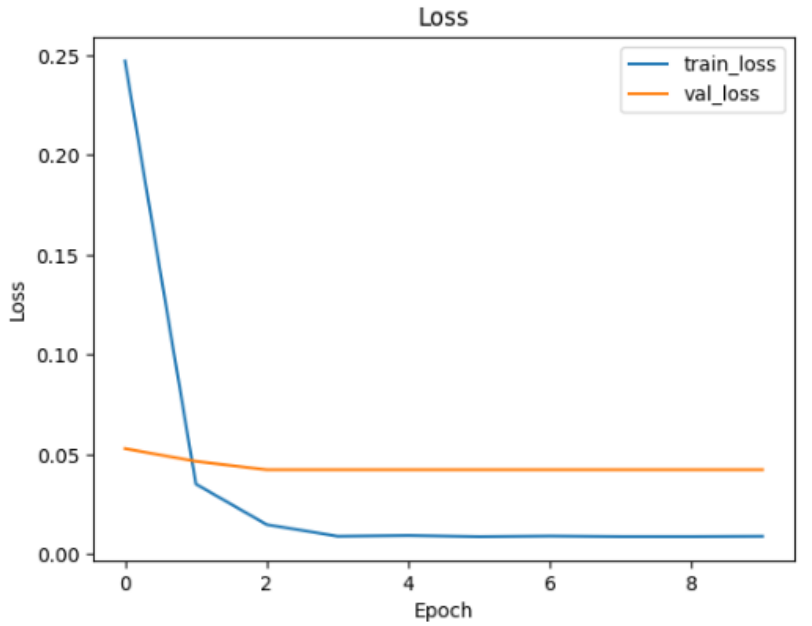
```python
    val_loss = 0.0
    model.eval()
    with torch.no_grad():
        total_correct = 0
        total_count = 0
        for batch in val_loader:
            batch = {k: v.to(device) for k, v in batch.items()}
            outputs = model(**batch)
            loss = outputs.loss
            val_loss += loss.item() * batch['input_ids'].shape[0]
            _, predicted = torch.max(outputs.logits, 1)
            total_correct += (predicted == batch['labels']).sum().item()
            total_count += len(batch['labels'])

        val_loss /= len(val_dataset)
        val_accuracy = total_correct / total_count
        val_loss_history.append(val_loss)
        val_acc_history.append(val_accuracy)
        print(f"Epoch {epoch} - Validation accuracy: {val_accuracy:.3f}")
```

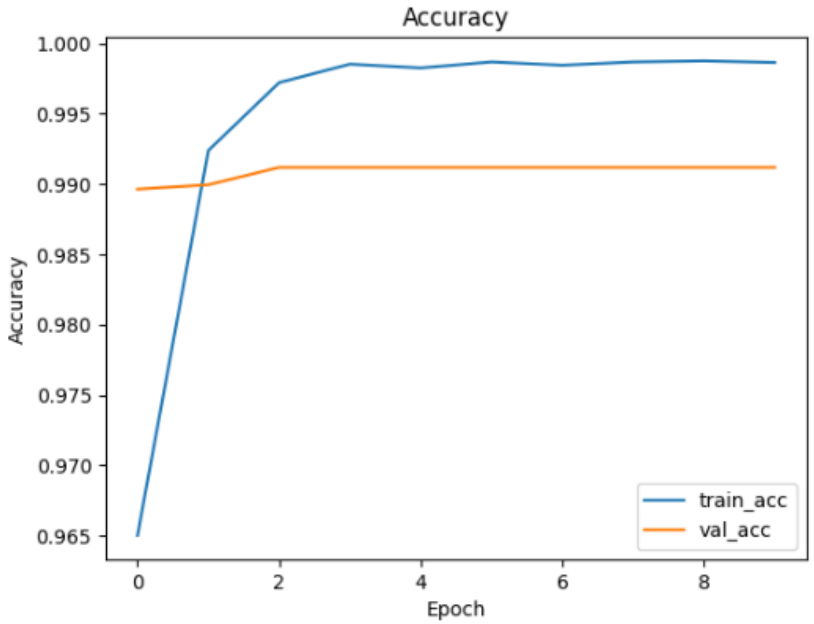
Experiments

```
Epoch 0 - Train loss: 0.190, Train accuracy: 0.971
Epoch 0 - Validation accuracy: 0.988
Epoch 1 - Train loss: 0.035, Train accuracy: 0.992
Epoch 1 - Validation accuracy: 0.990
Epoch 2 - Train loss: 0.015, Train accuracy: 0.996
Epoch 2 - Validation accuracy: 0.991
```

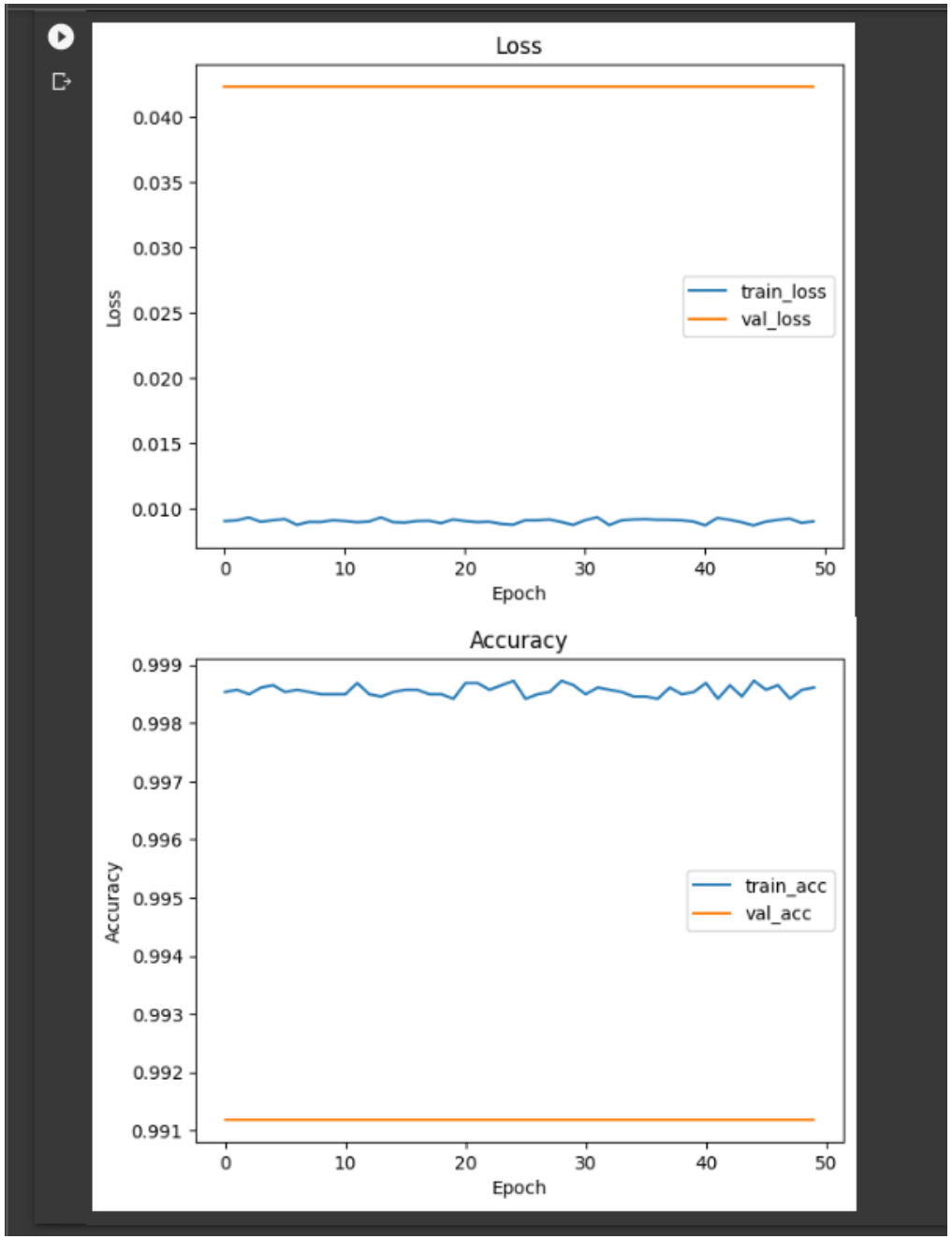
```
Epoch 10 - Train loss: 0.009, Train accuracy: 0.998
Epoch 10 - Validation accuracy: 0.991
Epoch 11 - Train loss: 0.009, Train accuracy: 0.999
Epoch 11 - Validation accuracy: 0.991
Epoch 12 - Train loss: 0.009, Train accuracy: 0.998
Epoch 12 - Validation accuracy: 0.991
Epoch 13 - Train loss: 0.009, Train accuracy: 0.998
Epoch 13 - Validation accuracy: 0.991
Epoch 14 - Train loss: 0.009, Train accuracy: 0.999
Epoch 14 - Validation accuracy: 0.991
Epoch 15 - Train loss: 0.009, Train accuracy: 0.999
Epoch 15 - Validation accuracy: 0.991
Epoch 16 - Train loss: 0.009, Train accuracy: 0.999
Epoch 16 - Validation accuracy: 0.991
Epoch 17 - Train loss: 0.009, Train accuracy: 0.998
Epoch 17 - Validation accuracy: 0.991
Epoch 18 - Train loss: 0.009, Train accuracy: 0.998
Epoch 18 - Validation accuracy: 0.991
Epoch 19 - Train loss: 0.009, Train accuracy: 0.998
Epoch 19 - Validation accuracy: 0.991
Epoch 20 - Train loss: 0.009, Train accuracy: 0.999
Epoch 20 - Validation accuracy: 0.991
Epoch 21 - Train loss: 0.009, Train accuracy: 0.999
Epoch 21 - Validation accuracy: 0.991
Epoch 22 - Train loss: 0.009, Train accuracy: 0.999
Epoch 22 - Validation accuracy: 0.991
Epoch 23 - Train loss: 0.009, Train accuracy: 0.999
Epoch 23 - Validation accuracy: 0.991
Epoch 24 - Train loss: 0.009, Train accuracy: 0.999
Epoch 24 - Validation accuracy: 0.991
Epoch 25 - Train loss: 0.009, Train accuracy: 0.998
Epoch 25 - Validation accuracy: 0.991
```

10 epochs





50 epochs



DEMO

- https://9337b92f13d65a2029.gradio.live/

Discussion

Potential Future Works

- Improve language detection and response features
  - Give more context
  - Respond in the language given as input
- Consider translation between languages
- Augment data for better performance
- Implement speech input, recognition and processing

Conclusion

- As we could tell our model wasn't too perfect, detection is pretty good, but for the question and response aspect, there is some work needed.
- Overall we enjoyed this project, it was super challenging, but we see ourselves continuing to improve it.

References

- https://insights2techinfo.com/generative-pre-trained-transformer/
- https://translate.google.com/?sl=en&tl=fr&text=I%27m%20hungry&op=translate
- https://openai.com
- https://arxiv.org/abs/1910.01108
- https://huggingface.co/docs/transformers/model_doc/distilbert
- https://gradio.app/docs/
- https://beebom.com/how-train-ai-chatbot-custom-knowledge-base-chatgpt-api/#:~:text=You%20can%20ask%20further%20questions,The%20possibilities%20are%20endless

THANK YOU!

Any questions?

Our model can answer it for you