SOC PROJECT

Language Detector And Translator

PROBLEM STATEMENT

- THE MAIN AIM OF THIS PROJECT IS TO CREATE A MODEL WHICH DETECTS THE LANGUAGE OF THE INPUT STRING.
- THE OTHER PROBLEM STATEMENT MAINLY INCLUDES BUILDING AN ENGLISH TO FRENCH LANGUAGE TRANSLATOR .
- THE MODEL SHOULD TAKE AN ENGLISH SENETENCE AS INPUT AND OUTPUT THE CORRESPONDING FRENCH TRANSLATION.

AI LANGUAGE DETECTOR

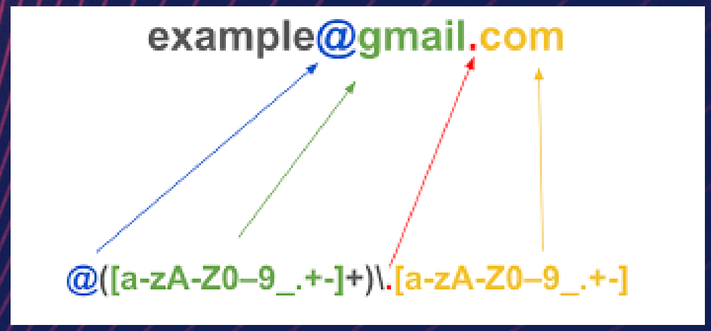
Overview

- THE LANGUAGE DETECTOR MAINLY DETECTS THE THE LANGUAGE OF THE STRING GIVEN TO IT AS INPUT BASED ON THE DATA IT IS TRAINED WITH.
- WE AR GIVEN A DATASET WITH TEO COLUMNS,ONE CONTAINING THE STRING AND THE OTHER CONTAINING THE LANGUAGE.
- IT IS TRAINED TO IDENTIFY ABOUT 22 LANGUAGES.

	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 เ...	Thai
3	விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...	Tamil
4	de spons behoort tot het geslacht haliclona en...	Dutch

PROCEDURE:

- FIRST THE DATA IS CLEANED USING THE REGEX LIBRARY

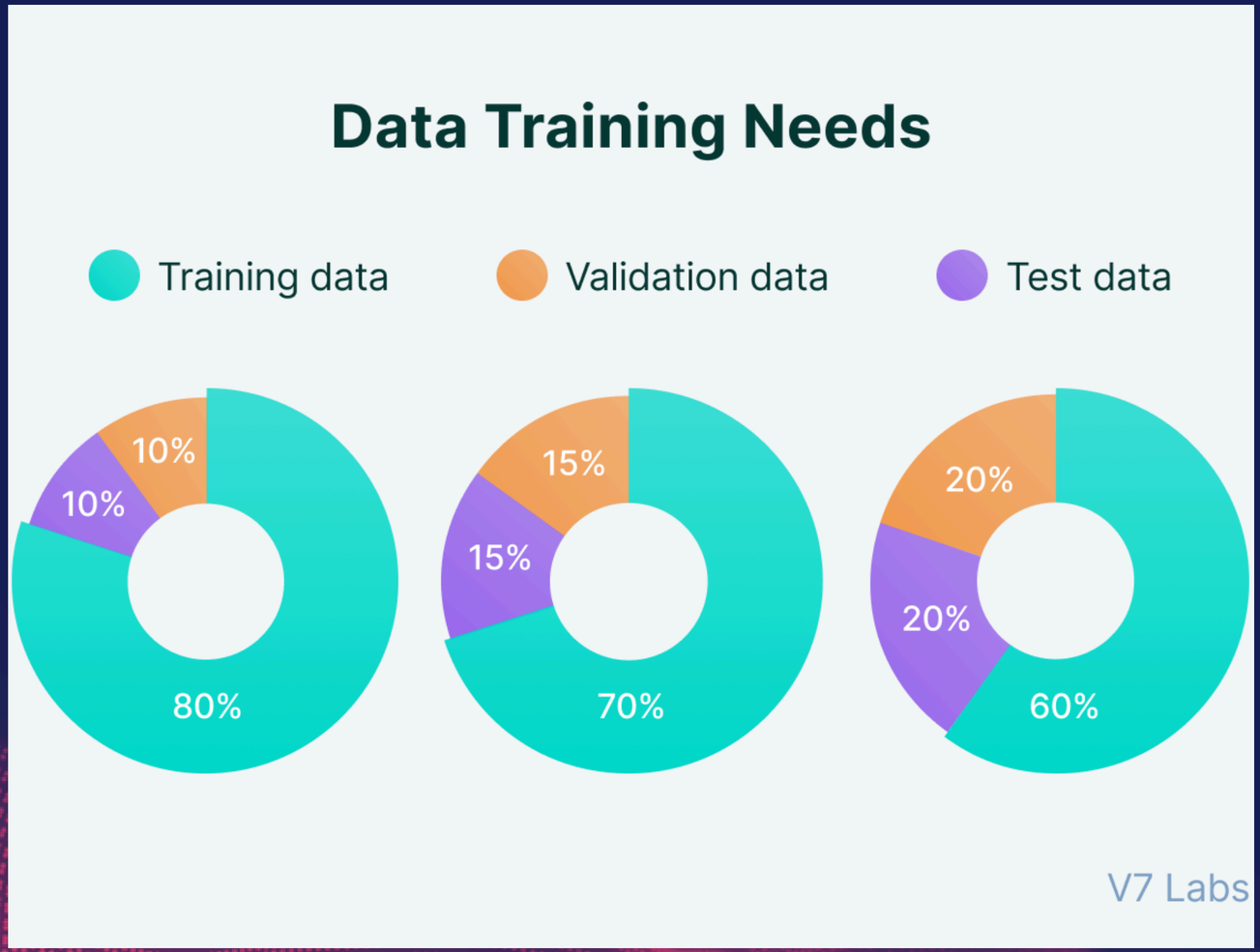


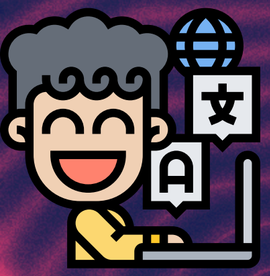
- THEN THE CLEANED DATA IS IS CONVERTED TO NUMERICAL SEQUENCES USING COUNTVECTORIZER

- THE LANGUAGES COLUMN IS ENCODED USING LABEL ENCODER.

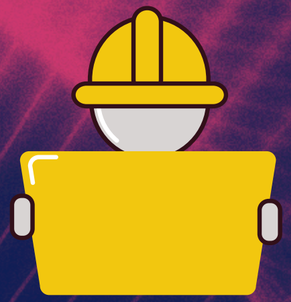
Formatting the num. sequences

- THE NUMERICAL SEQUENCES ARE THEN RESHAPED ACCORDINGLY.
- THE DATASET IS THEN SPLIT FOR TRAINING AND TESTING.
- THE TEST SIZE IS SET TO 0.2 PERCENT OF THE DATA SET AND IS ASLO USED FOR VALIDATION.

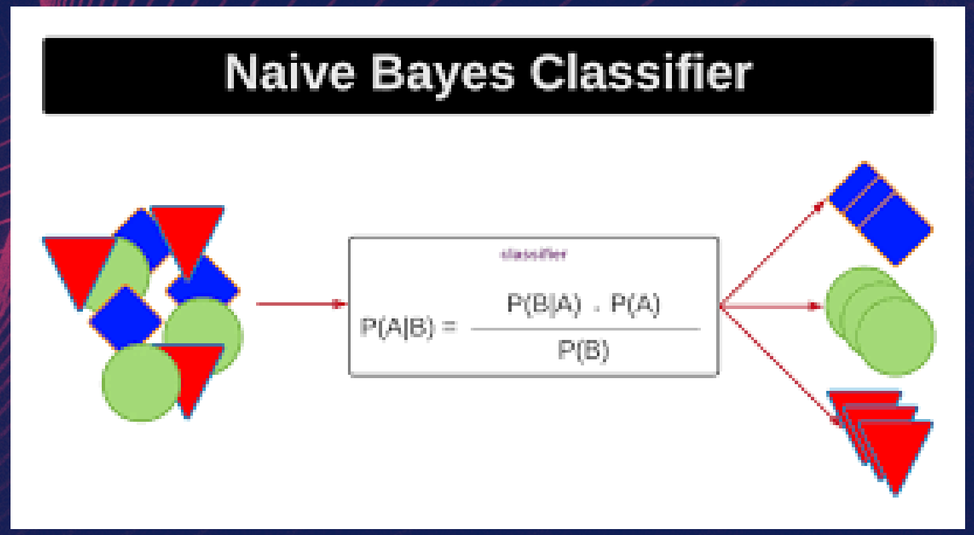




BUILDING THE MODEL



- THE MODEL WE ARE USING FOR THIS PROJECT IS NAIVE BAYES CLASSIFIER(MULTINOMIAL NB IN PARTICULAR)
- THE   MODEL IS TRAINED USING THE FIT FUNCTION AFTER CALLING AN INSTANCE OF MULTINOMIAL NB.



- THE MODEL IS THEN USED TO PREDICT ANY OTHER INPUT STRING.
-]THE MODEL IS THEN USED TO PREDICT THE TEST DATA AND THE PREDICTIONS ARE COMPARED WITH THE ACTUAL VALUES.
- THE OUTPUT IS IN THE FORM OF ARRAY WITH CORRESPONDING INDEX VALUES OF THE LANGUAGES AS DEFINED BY THE LABEL ENCODER.
- A DATAFRAME CAN BE USED TO COMPARE THE ACTUAL AND PREDICTED VALUES.

0.9197727272727273

[78]:

	y_pred	y_test
0	8	8
1	15	15
2	10	10
3	13	13
4	6	6
...	...	...
4395	2	2
4396	1	1
4397	3	3
4398	13	13
4399	2	2

4400 rows × 2 columns

```
def lang_det():

    mod=joblib.load('langmodel.pkl')
    a=input("give your string :")
    ap=preprocess(a)
    li=[]
    li.append(ap)
    t=cv.transform(li).toarray()
    t=t.tolist()
    fin=mod.predict(t)
    return le.classes_[fin[0]]
print(lang_det())
```

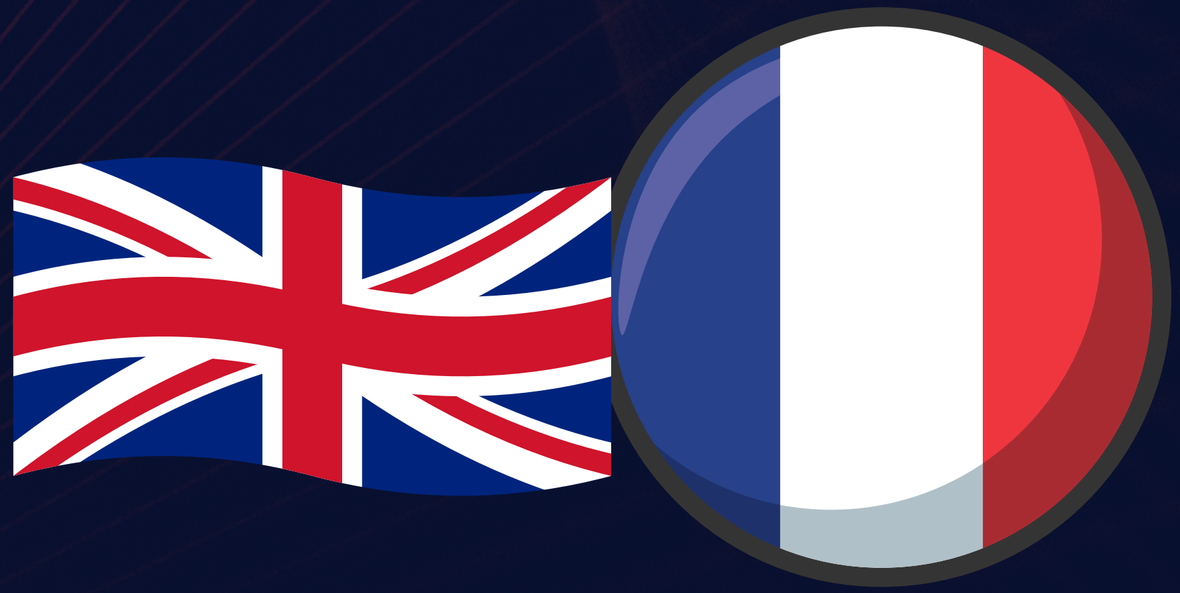
WE CAN THEN DEFINE A FUNCTION SUCH AS ABOVE WHICH GIVES THE FOLLOWING OUTPUT:

```
give your string : मेरे हाथ गन्दे है-
Hindi
```

PART 2: LANGUAGE TRANSLATOR

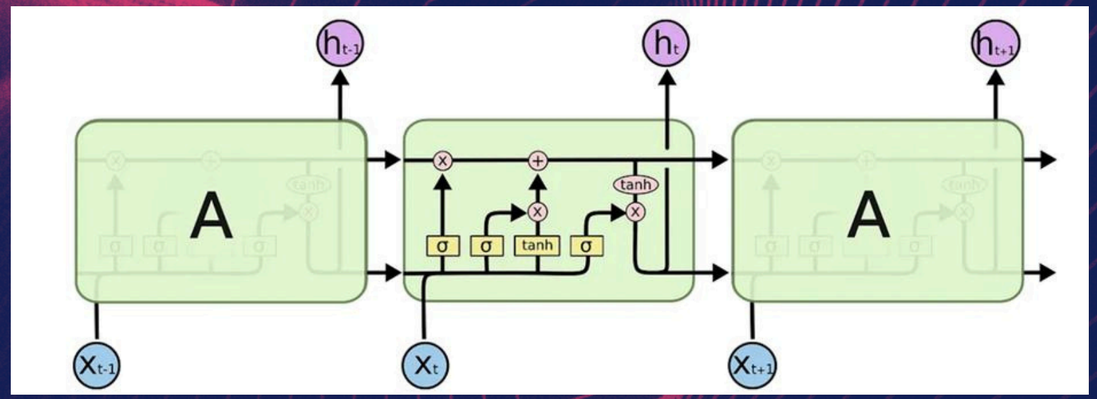
AIM: THE AIM OF THIS PROJECT INCLUDES CREATING AN ENGLISH TO FRENCH TRANSLATOR USING AN ENGLISH AND FRENCH DATASET.

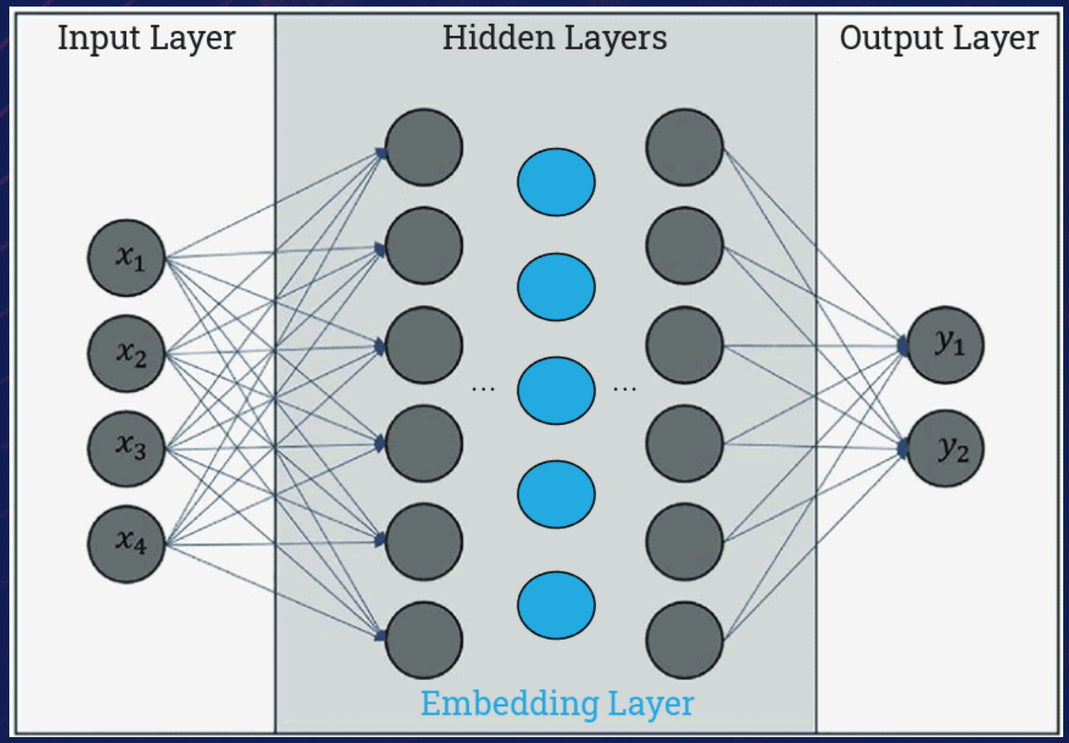
	English	French
0	Hi.	Salut!
1	Run!	Cours !
2	Run!	Courez !
3	Who?	Qui ?
4	Wow!	Ça alors !
5	Fire!	Au feu !
6	Help!	À l'aide !
7	Jump.	Saute.
8	Stop!	Ça suffit !
9	Stop!	Stop !



Procedure:

- The english and french data are encoded using tokenizer, one for english sentences and other for french sentences.
- Before tokenizing they are cleaned to remove any punctuations or unnecessary characters using regex.
- Then the encoded sequences are ready to be fed into the model.
- Here we are training the data using Embedding layer, LSTM,Repeat Vector layer and dense .





Following is the model summary

```
]: mod.summary()
```

Model: "sequential_24"

Layer (type)	Output Shape	Param #
embedding_26 (Embedding)	(None, 8, 100)	1,456,600
lstm_46 (LSTM)	(None, 256)	365,568
repeat_vector_19 (RepeatVector)	(None, 8, 256)	0
lstm_47 (LSTM)	(None, 8, 256)	525,312
time_distributed_4 (TimeDistributed)	(None, 8, 30819)	7,920,483

Total params: 30,803,891 (117.51 MB)

Trainable params: 10,267,963 (39.17 MB)

Non-trainable params: 0 (0.00 B)

Optimizer params: 20,535,928 (78.34 MB)

Predicting the translation

```
for word,index in fv.word_index.items():
    if index==1:
        print(f"{word}")
```

```
je
```

```
for word,index in ev.word_index.items():
    if index==2807:
        print(f"{word}")
```

```
hi
```

```
pred=mod.predict(np.array(x))
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 750ms/step
```

```
pred
```

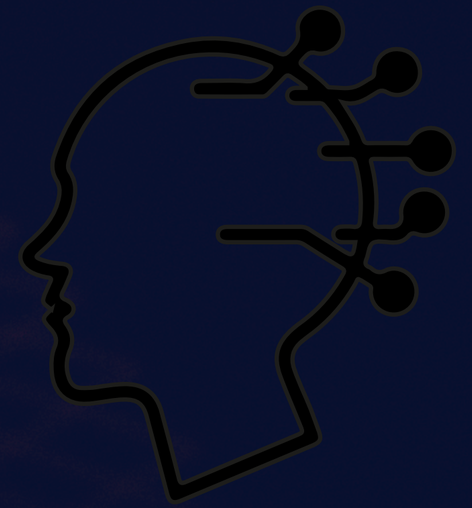
```
array([[[4.9198901e-05, 7.0211332e-05, 4.8575504e-05, ...,
         3.0956417e-05, 3.0954045e-05, 3.0957970e-05],
        [2.0355885e-04, 1.4913733e-04, 1.3172172e-04, ...,
         2.6288213e-05, 2.6001251e-05, 2.6245511e-05],
        [2.0898687e-02, 4.3450386e-04, 1.6597810e-03, ...,
         6.3513276e-06, 6.0216307e-06, 6.2967829e-06],
        ...,
        [8.7355983e-01, 2.3348324e-04, 3.8723259e-03, ...,
         8.0126683e-10, 6.9860140e-10, 7.9227774e-10],
        [8.7480205e-01, 2.3168020e-04, 3.8576983e-03, ...,
         7.7459517e-10, 6.7519390e-10, 7.6597539e-10],
        [8.7496841e-01, 2.3143430e-04, 3.8557192e-03, ...,
         7.7105555e-10, 6.7209049e-10, 7.6248097e-10]]], dtype=float32)
```

- After fitting the model wecan now predict the french translation for the english sentence or word
- The output is of the form as shown in right side and the translation is found by using argmax function of numpy.(It gives the index of highest probability.
- The translation can then be found from the word index of french vocabulary.

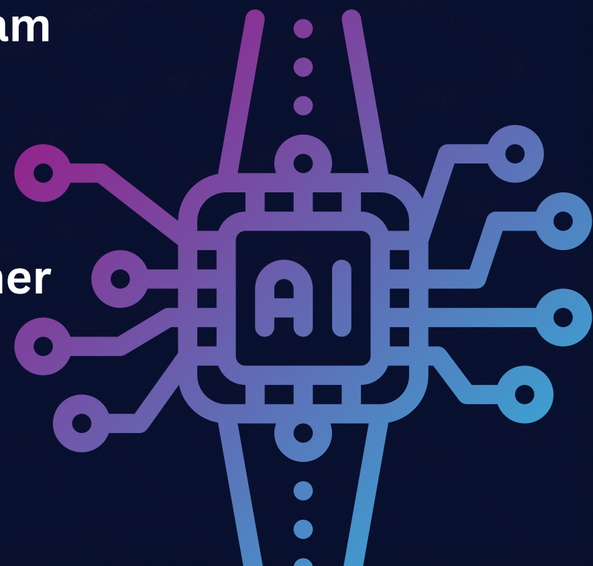
Note: The model developed here is of low accuracy.

So the tranlsations are not exact .

KEY LEARNINGS



- The SOC project has been a wonderful experience and I was able to learn a lot of things.
- This project helped me delve deep into world of machine learning and neural networks.
- I learnt about different neural networks like RNN, CNN,LSTM,embedding ,etc and their working.
- Many useful modules for NLP like regex, spacy,nltk,etc which are used for text processing.
- I also tried some projects from youtube videos such as email spam and ham classification, Sentiment analysis, Digit Finder and some others.



Overall I found this journey very interesting and looking forward to take many other projects in the upcoming future.

Thank you!