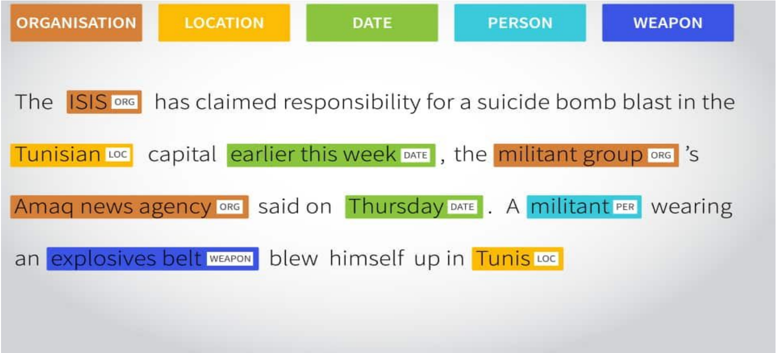
Named Entity Recognition - UN Speech Transcripts

Aishwarya Sarkar (as99646), Aniket Patil (aap3788),
Anudeep Kumar Akkana (aa92799), Pratik Gawli (pbg397)

What is Named Entity Recognition?

NER is a task of extracting information from the sequence of words and sentences and classifying them into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

It is a subtask of information extraction.



Uses of Named Entity Recognition

1. Classifying content for news providers
2. Powering Content Recommendations
3. Entity Detection in Research Papers

UN NER Dataset Description

- The dataset consists of speeches given at the United Nations General Assembly from 1993-2016 scraped from the website and then parsed
- There are a total of 70 labeled documents consisting of transcribed speeches which 50 in the training and 20 in the test data
- More than 50,000 tokens in the test data were manually tagged for Named Entity Recognition (O - Not a Named Entity; I-PER - Person; I-ORG - Organization; I-LOC - Location; I-MISC - Other Named Entity)

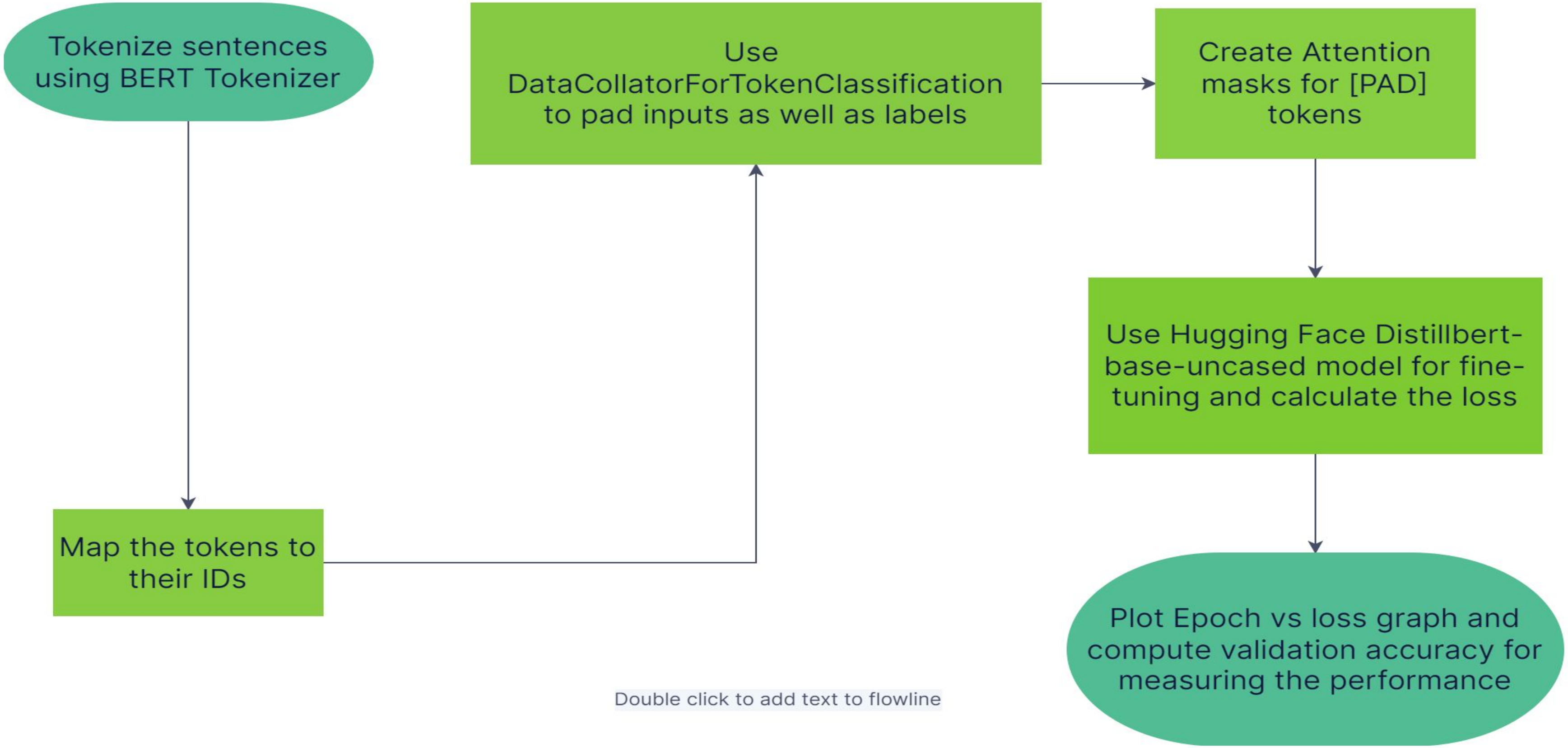
Hugging Face transformers

Hugging Face standardizes all the steps involved in training and using a language model. It has an API that allows easy access to pretrained models, datasets and tokenizing steps.

Model in use: ***DistilBERT*** *(distilbert-base-uncased)*
DistilBERT is a transformers model, smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion, using the BERT base model as a teacher.This model is uncased: it does not make a difference between english and English.

In DistilBERT, the size of a BERT model is reduced by 40% via knowledge distillation during the pre-training phase while 97% of its language understanding abilities is retained. It is also 60% faster.

Process Flow Chart



Tokenization and Prep-processing

- Extract tokens and map tags from training dataset
  - Used BIO scheme of tagging
    - B: Token is start of a named entity (Used only when entity has multiple tokens)
    - I: Token is inside a named entity
    - O: Token is not a named entity

tokens	ner_tags
[This, new, session, of, the, Assembly, of, th...	[O, O, O, O, O, I-ORG, O, O, I-ORG, I-ORG, O, ...
[These, are, certainly, critical, times, that,...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ...

- Above dataframe is mapped to **distilbert-base-uncased tokenizer** which creates **attention masks** to be used for fine-tuning in next steps

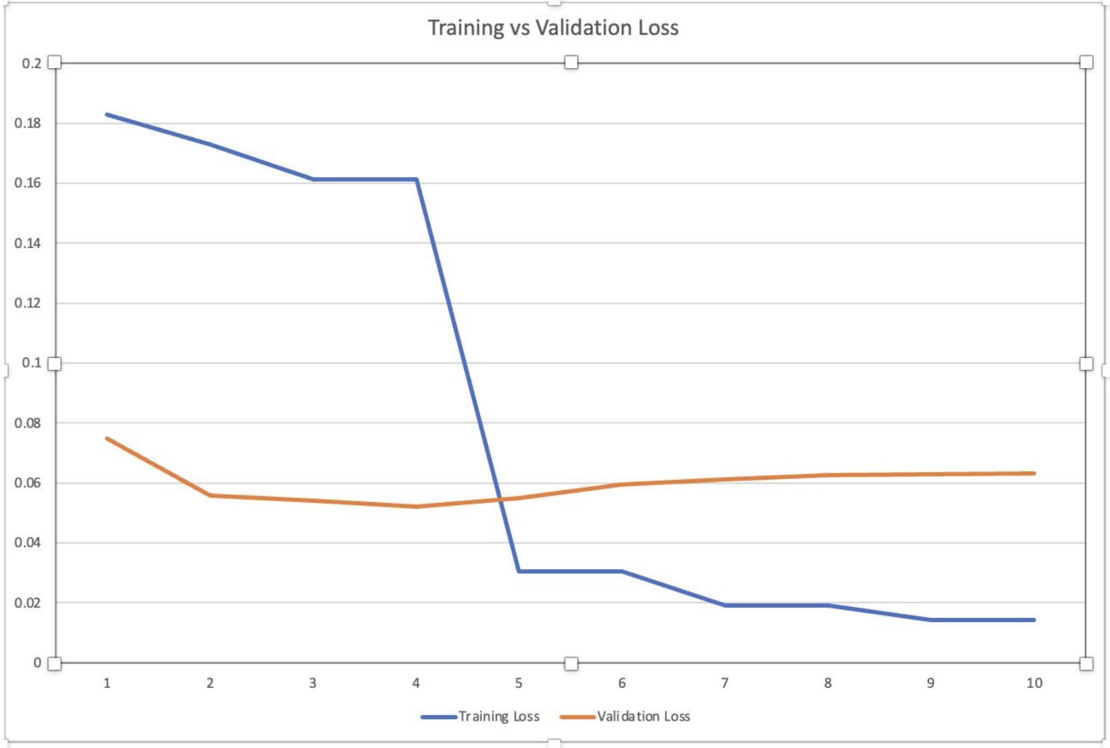
Fine-tuning & Hyperparameter opt.

- Used distilbert-case-uncased model using “AutoModelTokenClassification” class
- Used “DataCollatorForTokenClassification” for collating data for NER task
- Used “Trainer” class for fine-tuning with UN transcripts
- Optimize hyper-parameters using Ray Tune
    - Batchsize, learning rate, epochs
- Save optimized fine-tuned model
- Load model for tokenization followed by entity recognition

Results

- *A named entity is correct only if it is an exact match of the corresponding entity in the data file*
- *Metrics used to evaluate the NER model: Precision, Recall, F1 Score and Accuracy.*
- ***Accuracy**: Accuracy is the proportion of correct predictions among the total number of cases processed.*
- ***Precision**: Percentage of named entities found by the learning system that are correct.*
- ***Recall** : Percentage of named entities present in the corpus that are found by the system.*
- ***F1 Score :** The harmonic mean of the precision and recall.*
- *Used the **load_metric** function from the datasets library in Hugging Face to load the **seqeval** metric*

Results



Hyperparameters:

10 epochs
Batch size = 16
Learning rate=1e-5

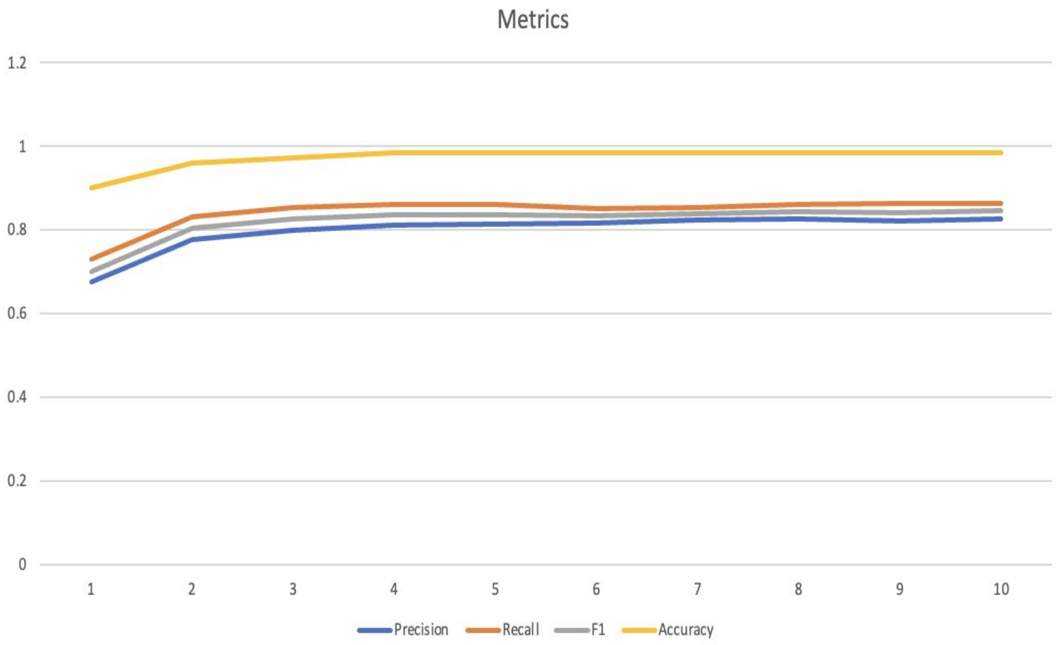
Loss for training and validation sets:

Both training and validation losses are high at epoch 1(High Bias, Low variance)

Training loss decreases with increasing number of epochs, as expected

Validation loss lowest at the 4th Epoch, increases after that due to overfitting. (Low Bias, High Variance)

Results



Accuracy improves but is almost close to 1 throughout all the epochs

Precision, Recall and F1 score better metrics to evaluate NER model

Precision: Correctly identified Named Entities from the training set increases with number of epochs. Maximum: 0.8268

Recall: Percentage of named entities present in the corpus that are found follows a similar trajectory as Precision.0.864

Results

Sample prediction

"Congratulations to Mr. Johnson on his assumption of the Presidency of the General Assembly in Switzerland at its sixty-sixth session."

Without fine-tuning

	words	ner
0	[CLS]	I-MISC
1	congratulations	B-MISC
2	to	B-ORG
3	mr	B-ORG
4	.	B-ORG
5	johnson	B-ORG
6	on	B-MISC
7	his	B-ORG
8	assumption	B-ORG
9	of	B-ORG
10	the	B-ORG
11	presidency	B-PER
12	of	B-MISC
13	the	B-ORG

	words	ner
14	general	B-PER
15	assembly	B-PER
16	in	B-MISC
17	switzerland	I-PER
18	at	B-MISC
19	its	B-ORG
20	sixty	B-LOC
21	-	B-LOC
22	sixth	B-PER
23	session	B-PER
24	.	B-ORG
25	[SEP]	B-ORG

With fine-tuning

	words	ner
0	[CLS]	O
1	congratulations	O
2	to	O
3	mr	O
4	.	O
5	johnson	I-PER
6	on	O
7	his	O
8	assumption	O
9	of	O
10	the	O
11	presidency	O
12	of	O
13	the	O

	words	ner
14	general	I-ORG
15	assembly	I-ORG
16	in	O
17	switzerland	I-LOC
18	at	O
19	its	O
20	sixty	O
21	-	O
22	sixth	O
23	session	O
24	.	O
25	[SEP]	I-MISC

Conclusion

Fine-tuned the Hugging Face NER model successfully with UN transcripts data to perform NER on the generated transcripts.

This model can be used in assisting UN speech transcript analysis in order to track the trends of which entities are most discussed during which time of the year to generate insights

Fine tuning can be done using other datasets as well such as sensitive information labels which can be used to censor confidential documents

Metrics such as Recall and Precision help evaluate the fine-tuning of the model better than the accuracy.

Thank you!