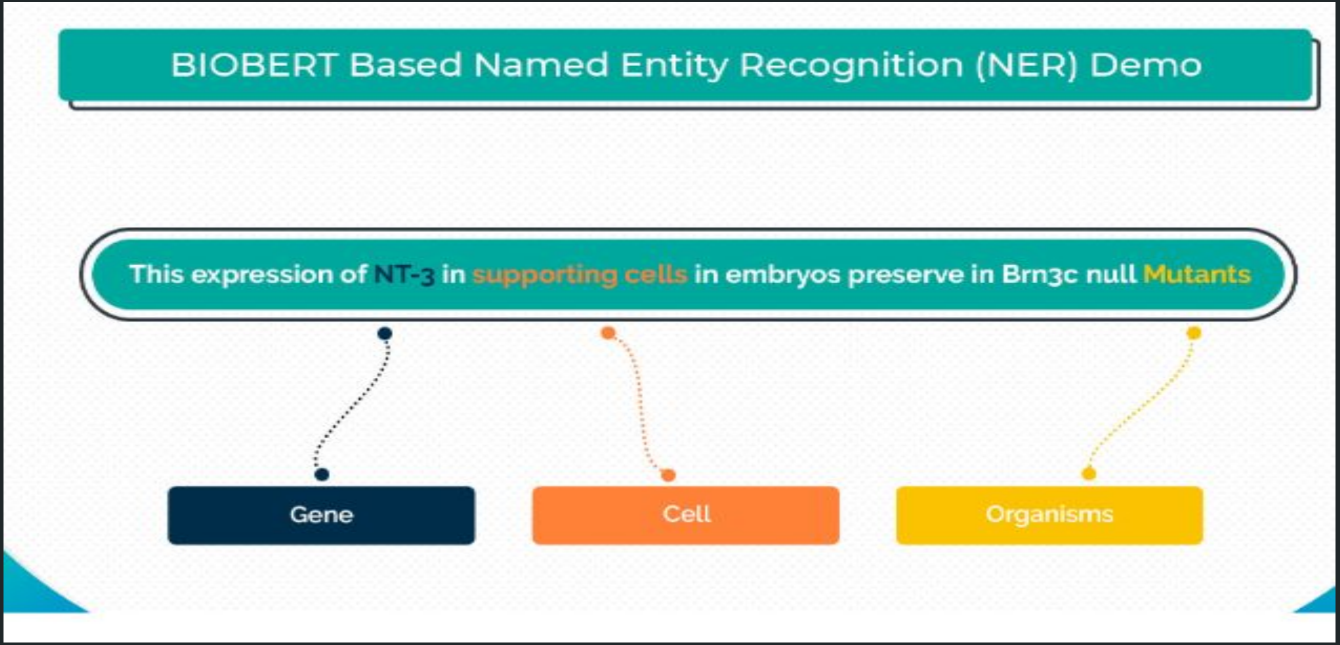
BIOMEDICAL NER

INTRODUCTION

- Recognition of biomedical entities in biomedical research papers is a challenging task.
- Biomedical NER aims to recognize biomedical entities - chemicals, diseases, proteins and genes in a given text.
- *Named entity recognition* involves recognizing numerous domain-specific proper nouns in a biomedical corpus.
- BioBERT directly learns WordPiece embeddings during pre-training and fine-tuning.
- For the evaluation metrics of NER, we used F1 score.

NAMED ENTITY RECOGNITION



BIOBERT Based Named Entity Recognition (NER) Demo

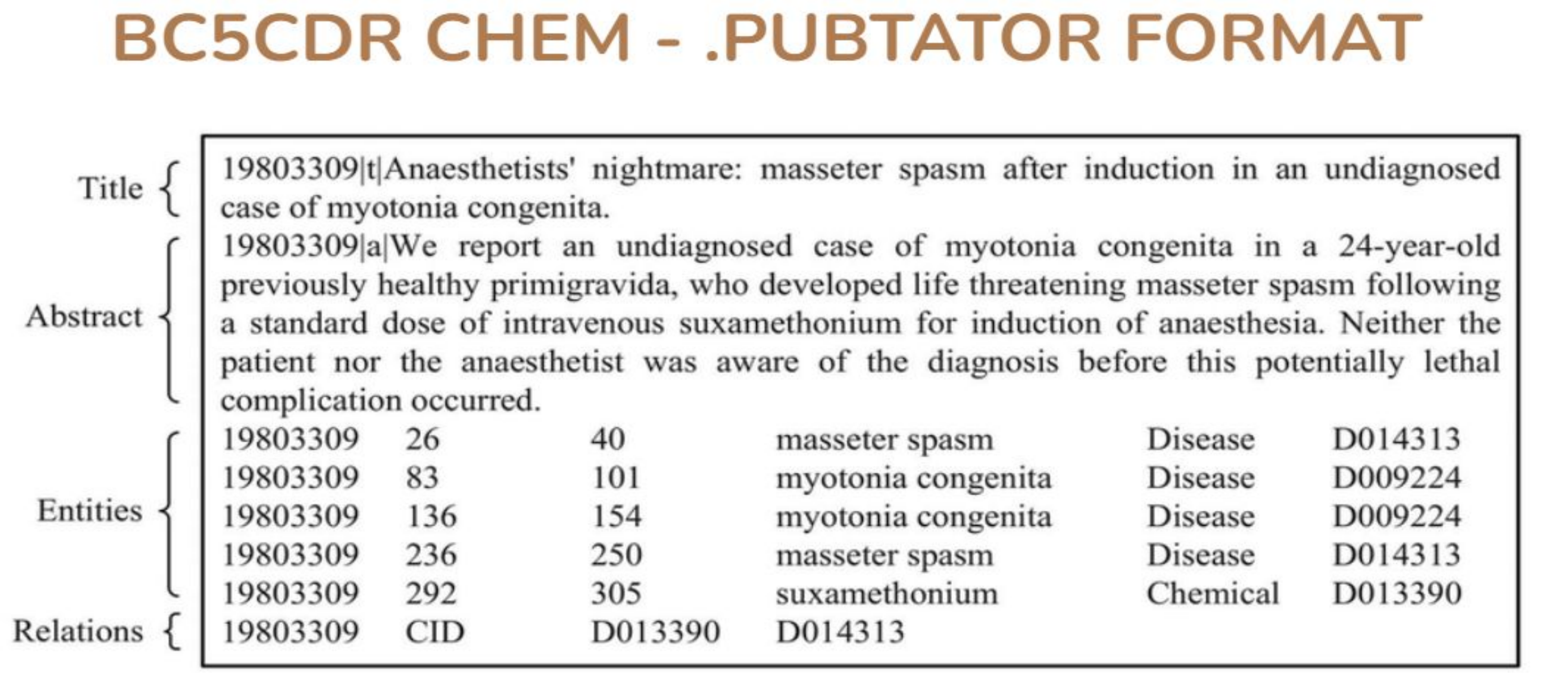
This expression of NT-3 in supporting cells in embryos preserve in Brn3c null Mutants

DATA USED

- BC5CDR - 1500 PubMed articles
  - 4409 chemicals
  - 5818 diseases
  - 3116 chemical-disease interactions

- CHEMPROT - 1820 PubMed articles
  - Chemical-protein interactions annotated by domain experts
  - Used in the BioCreative VI text mining chemical-protein interactions shared task.
  - Contains entities such as Chemical , GENE

DATA FORMATS



BC5CDR CHEM - .PUBTATOR FORMAT

Title

```
19803309|t|Anaesthetists' nightmare: masseter spasm after induction in an undiagnosed case of myotonia congenita.
```

Abstract

```
19803309|a|We report an undiagnosed case of myotonia congenita in a 24-year-old previously healthy primigravida, who developed life threatening masseter spasm following a standard dose of intravenous suxamethonium for induction of anaesthesia. Neither the patient nor the anaesthetist was aware of the diagnosis before this potentially lethal complication occurred.
```

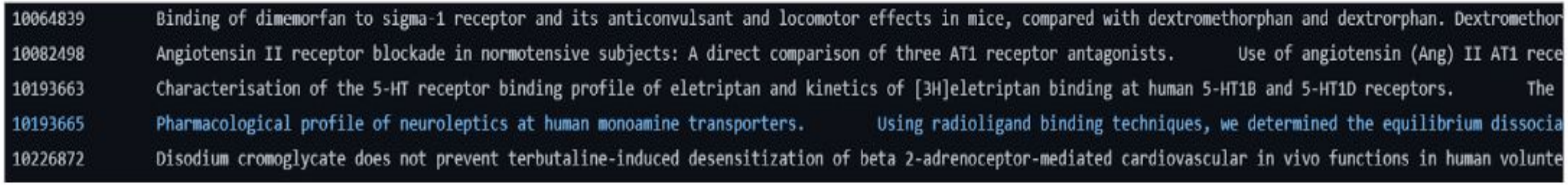
Entities

```
19803309    26     40         masseter spasm        Disease     D014313
19803309    83     101        myotonia congenita    Disease     D009224
19803309    136    154        myotonia congenita    Disease     D009224
19803309    236    250        masseter spasm        Disease     D014313
19803309    292    305        suxamethonium         Chemical    D013390
```

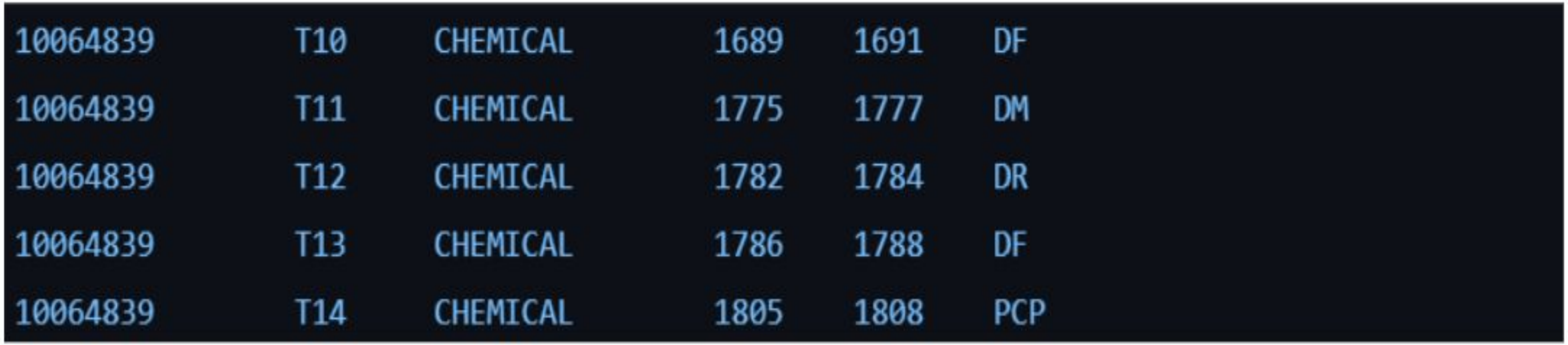
Relations

```
19803309    CID    D013390    D014313
```

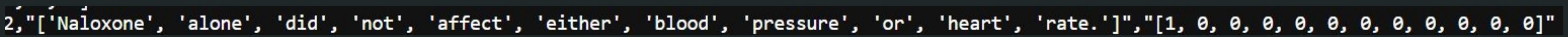
CHEMPROT - .TSV FORMAT



```
10064839	Binding of dimemorfan to sigma-1 receptor and its anticonvulsant and locomotor effects in mice, compared with dextromethorphan and dextrorphan. Dextromethor
10082498	Angiotensin II receptor blockade in normotensive subjects: A direct comparison of three AT1 receptor antagonists.	Use of angiotensin (Ang) II AT1 rece
10193663	Characterisation of the 5-HT receptor binding profile of eletriptan and kinetics of [3H]eletriptan binding at human 5-HT1B and 5-HT1D receptors.	The
10193665	Pharmacological profile of neuroleptics at human monoamine transporters.	Using radioligand binding techniques, we determined the equilibrium dissocia
10226872	Disodium cromoglycate does not prevent terbutaline-induced desensitization of beta 2-adrenoceptor-mediated cardiovascular in vivo functions in human volunte
```



```
10064839	T10	CHEMICAL	1689	1691	DF
10064839	T11	CHEMICAL	1775	1777	DM
10064839	T12	CHEMICAL	1782	1784	DR
10064839	T13	CHEMICAL	1786	1788	DF
10064839	T14	CHEMICAL	1805	1808	PCP
```



```
2,"['Naloxone', 'alone', 'did', 'not', 'affect', 'either', 'blood', 'pressure', 'or', 'heart', 'rate.']","[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]"
```

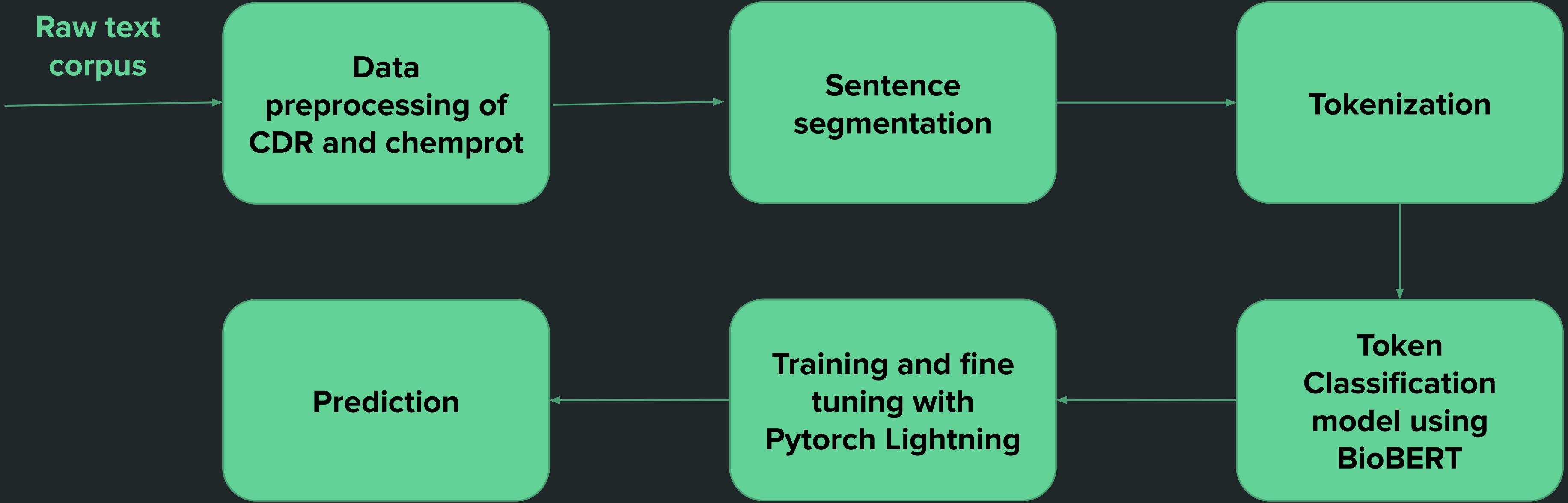
BIOBERT - Bidirectional Encoder Representations from Transformers for Biomedical Text Mining

- Directly applying the advancements in NLP to biomedical text mining often yields unsatisfactory results due to a word distribution shift from general domain corpora to biomedical corpora.

- BioBERT which is a domain-specific language representation model pre-trained on large-scale biomedical corpora.

- BioBERT largely outperforms BERT and previous state-of art models

- BioBERT significantly outperforms them on biomedical text mining tasks such as NER , RE and QA



Raw text corpus

Data preprocessing of CDR and chemprot

Sentence segmentation

Tokenization

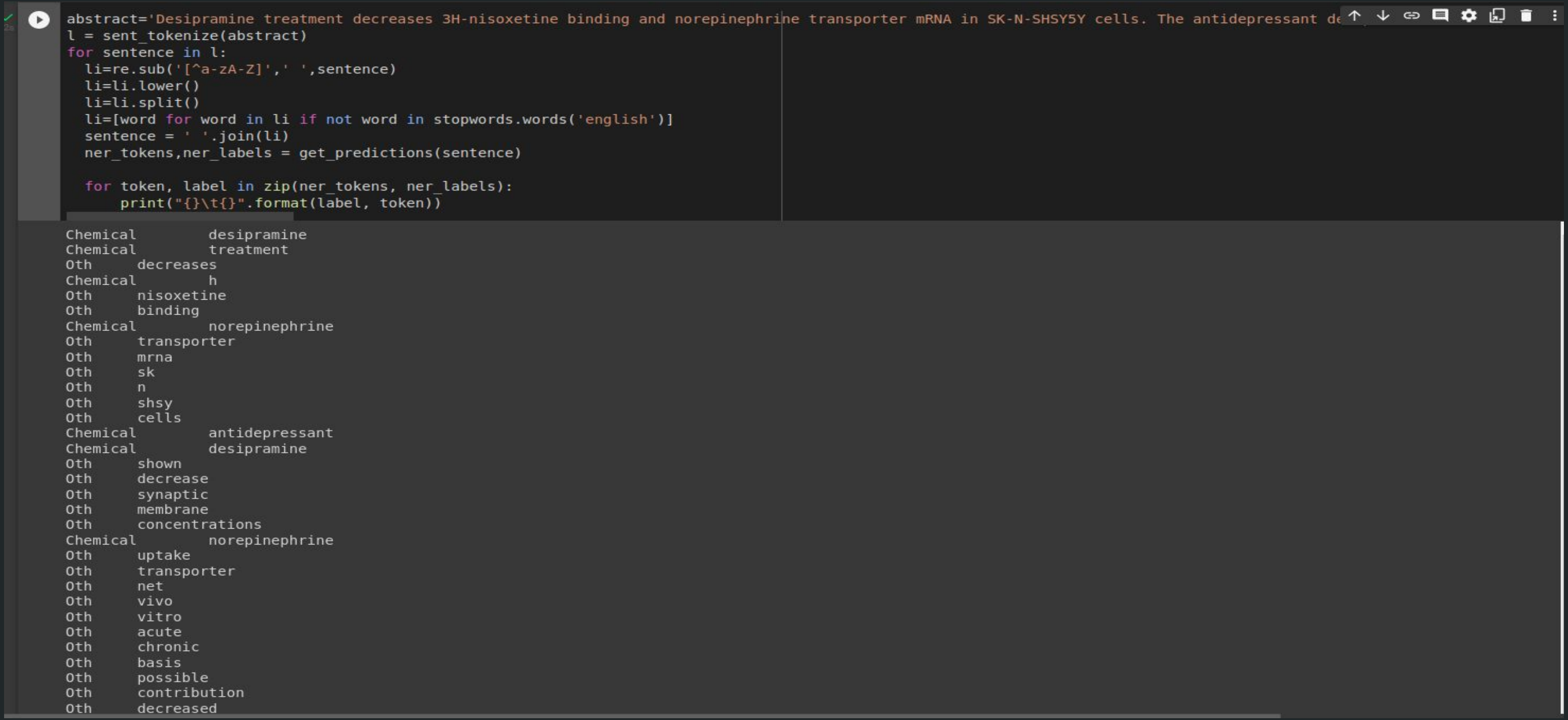
Token Classification model using BioBERT

Training and fine tuning with Pytorch Lightning

Prediction

MODEL PIPELINE

PREDICTIONS

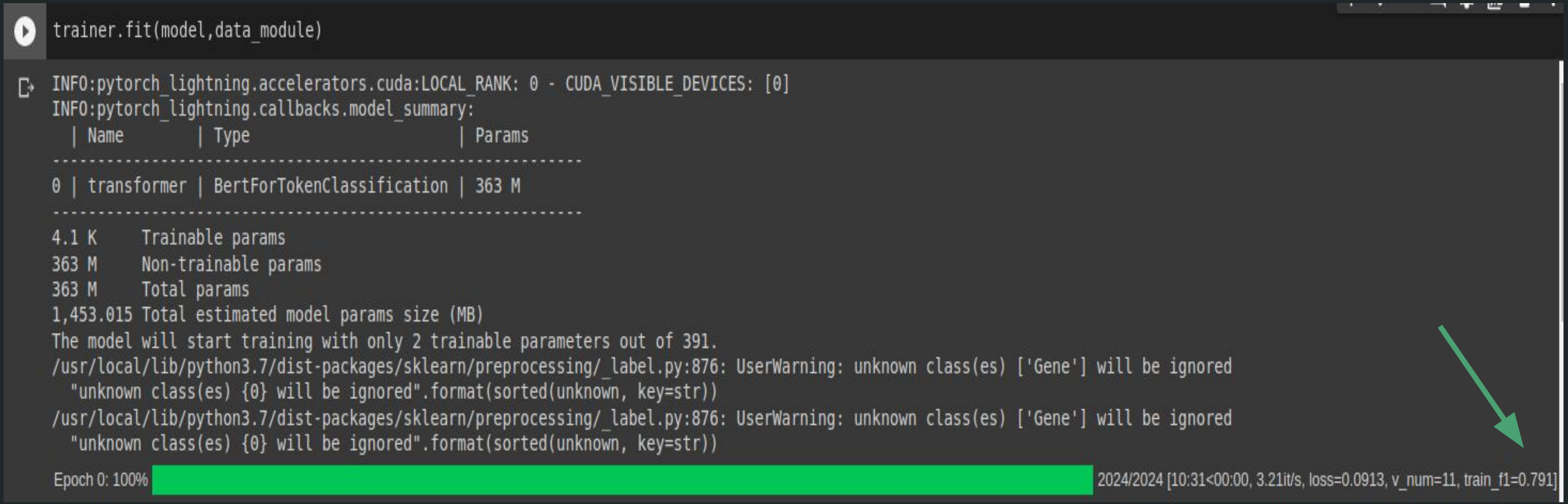


```
abstract='Desipramine treatment decreases 3H-nisoxetine binding and norepinephrine transporter mRNA in SK-N-SHSY5Y cells. The antidepressant de
l = sent_tokenize(abstract)
for sentence in l:
  li=re.sub('[^a-zA-Z]',' ',sentence)
  li=li.lower()
  li=li.split()
  li=[word for word in li if not word in stopwords.words('english')]
  sentence = ' '.join(li)
  ner_tokens,ner_labels = get_predictions(sentence)

  for token, label in zip(ner_tokens, ner_labels):
      print("{}\t{}".format(label, token))
```

```
Chemical        desipramine
Chemical        treatment
Oth     decreases
Chemical        h
Oth     nisoxetine
Oth     binding
Chemical        norepinephrine
Oth     transporter
Oth     mrna
Oth     sk
Oth     n
Oth     shsy
Oth     cells
Chemical        antidepressant
Chemical        desipramine
Oth     shown
Oth     decrease
Oth     synaptic
Oth     membrane
Oth     concentrations
Chemical        norepinephrine
Oth     uptake
Oth     transporter
Oth     net
Oth     vivo
Oth     vitro
Oth     acute
Oth     chronic
Oth     basis
Oth     possible
Oth     contribution
Oth     decreased
```

RESULTS



```
trainer.fit(model,data_module)
```

```
INFO:pytorch_lightning.accelerators.cuda:LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES: [0]
INFO:pytorch_lightning.callbacks.model_summary:
  | Name        | Type                       | Params
-----------------------------------------------------------
0 | transformer | BertForTokenClassification | 363 M
-----------------------------------------------------------
4.1 K     Trainable params
363 M     Non-trainable params
363 M     Total params
1,453.015 Total estimated model params size (MB)
The model will start training with only 2 trainable parameters out of 391.
/usr/local/lib/python3.7/dist-packages/sklearn/preprocessing/_label.py:876: UserWarning: unknown class(es) ['Gene'] will be ignored
  "unknown class(es) {0} will be ignored".format(sorted(unknown, key=str))
/usr/local/lib/python3.7/dist-packages/sklearn/preprocessing/_label.py:876: UserWarning: unknown class(es) ['Gene'] will be ignored
  "unknown class(es) {0} will be ignored".format(sorted(unknown, key=str))
Epoch 0: 100% 2024/2024 [10:31<00:00, 3.21it/s, loss=0.0913, v_num=11, train_f1=0.791]
```

The obtained F1 score for one epoch is approximately 80%

CONCLUSION

- We have successfully built a generalised model with approximately 80% F1 score for the prediction of chemical, disease and gene in the research articles.

- We can attain more accuracy if we train the model on more epochs with high GPU acceleration.

- We have fine-tuned the model with the help of pytorch lightning

THANK YOU