NER-using BERT

ELL-881 Assignment-3 Report

Vaibhav Seth
2021MT10236

Data Cleaning

There were certain examples in the data where the number of tokens in the text didn't match the number of NER-labels. Those examples were removed from the DataFrame.

Data Pre-processing

Because of constraints over the GPU memory, we consider only those input texts whose tokenized versions (without [CLS] and [SEP]) were of length <= 128.

A set of unique labels were generated by iterating over df['labels']. These labels were converted to ids (0-17) for and also, -100 was assigned as the label for pad token.

The BERT tokenizer might split a word into multiple sub-word tokens, hence resulting in an apparent change in the input length. To counter this effect on the output label side, the io_align() function was used. The function tokenizes each of the word into their corresponding sub-words and assigns the label of the original word to all of them. This way we get same length input-output pairs.

For Batch purposes all the sentences were padded with max length = (128), and the pad tokens were assigned with a label of -100.

A torch.utils.data.Dataset class was created to provide the data to the DataLoader for BERT.

MODEL

We use the BertForTokenClassification model from the HuggingFace library. It is a standard encoder-decoder model with a classification head on top. The weights were initialized using the pre-trained 'bert-base-cased' model. We use the cased version as it would better capture the structure of the sentences and tags.

The model takes in 'input_ids', 'attention_mask' and 'labels' as inputs and returns the loss between predicted labels and given labels, and the predicted outputs (in form of the last hidden state).

We take the output of the last hidden state and apply softmax to each all the tokens and take their argmax to get the predicted label.

In the accuracy calculation we do not consider those tokens whose ground truth labels are -100 ([PAD]) or 'O'.

Training setup

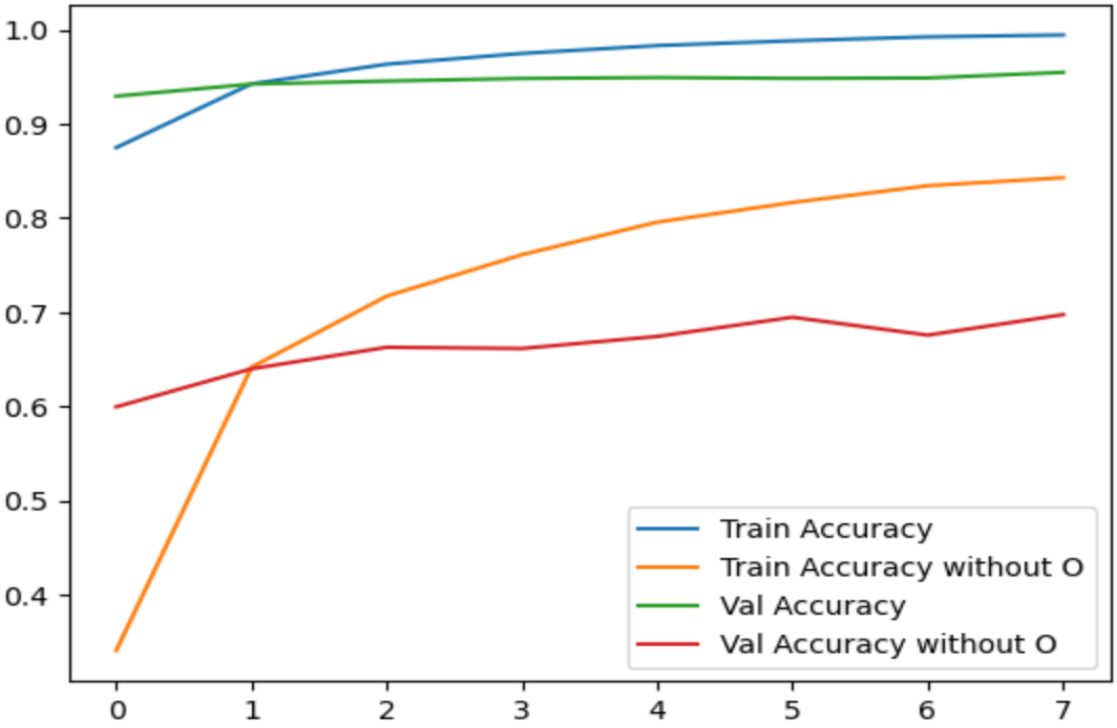
The model was trained using **AdamW** optimizer with a learning rate of 3e-5. The models were trained over 8 epochs, with a batch size of 16.

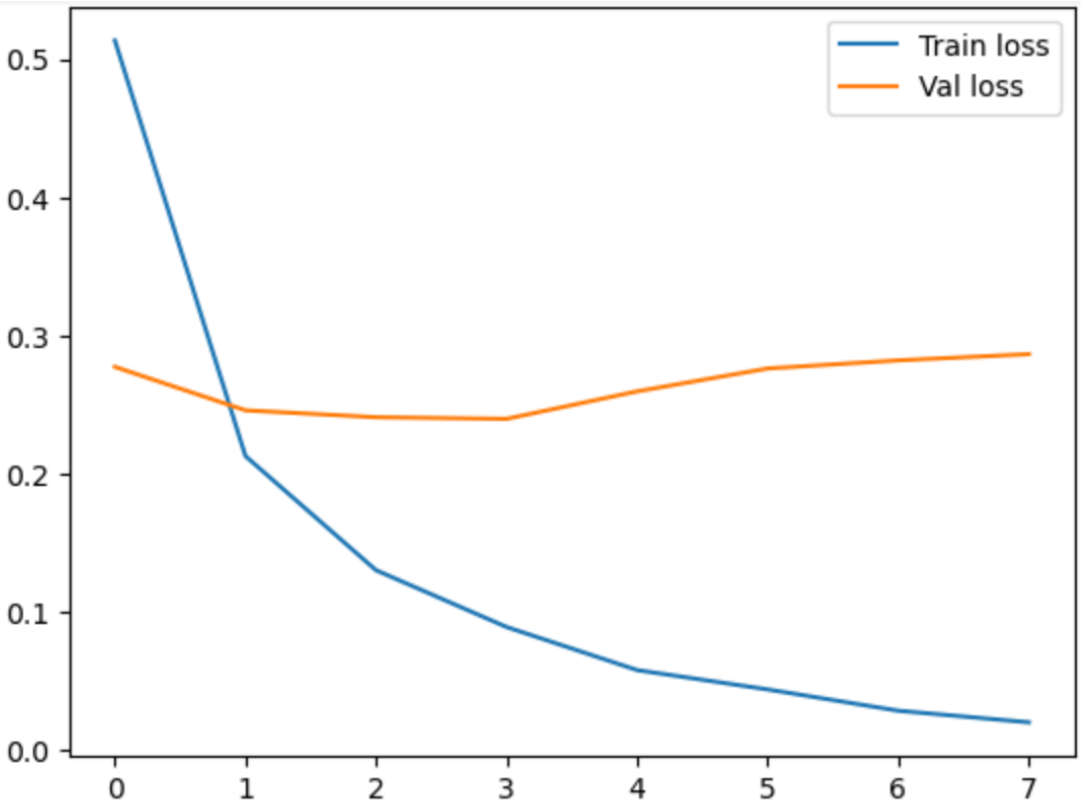
The number of examples used for training and validation vary as 1000,2000,5000,10000.

Results

All the examples have been split in the ratio of 8:1:1

1) Using 2000 examples





Test Set Examples

European observers and U . S . officials also have strongly criticized the election proc ess .
B-gpe O O B-geo B-geo B-geo B-geo O O O O O O O O O <- Predicted
O O O B-geo B-geo B-geo B-geo O O O O O O O O O <- Actual
Accuracy = tensor(1.)

China has rejected past over ##tures made by Mr . Chen , and has refused to speak w ith the Taiwanese leader until he agrees that Taiwan is an ins ##ep ##ara ##ble part o f China .
B-geo O O O O O O O B-per B-per I-per O O O O O O O O B-gpe O O O O O B-geo O O O O O O O O B-geo O <- Predicted
B-geo O O O O O O O B-per B-per I-per O O O O O O O O B-gpe O O O O O B-geo O O O O O O O O B-geo O <- Actual
Accuracy = tensor(1.)

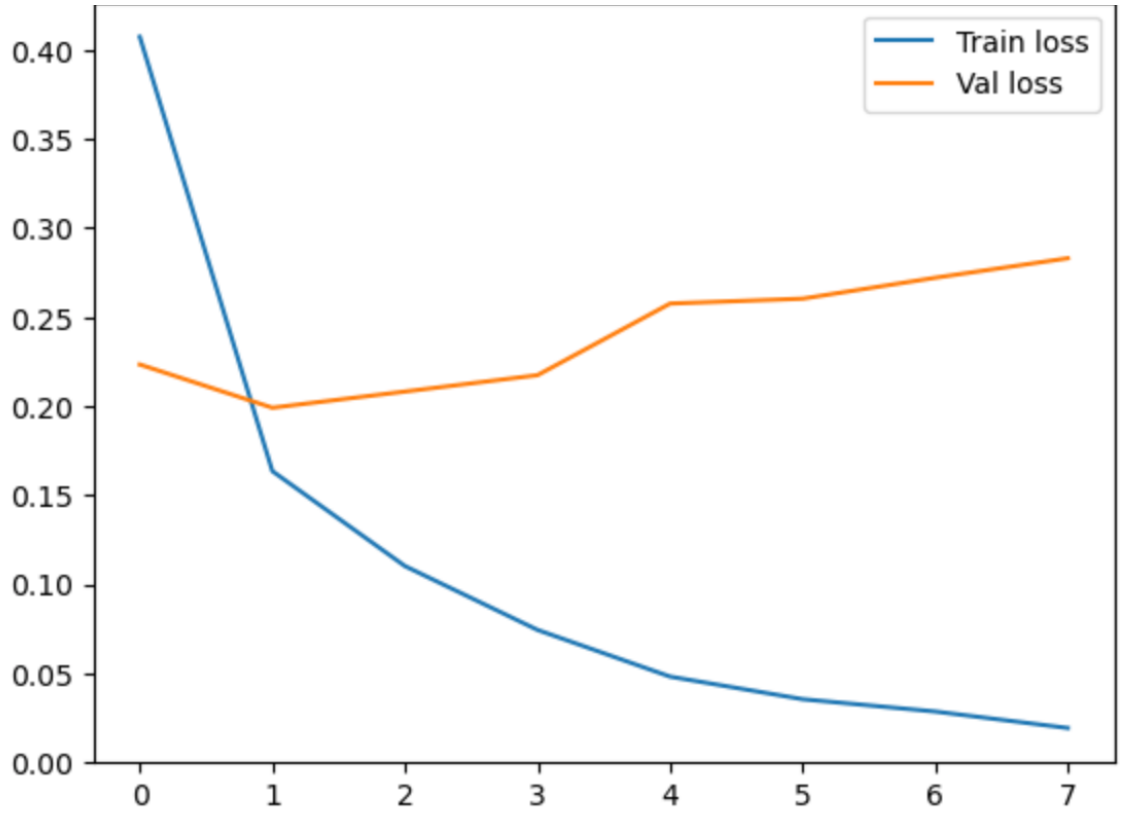
The head of Cuba ' s parliament has offered to support Iran in its fight to develop nuc lear energy within Iranian borders .
O O O B-geo O O O O O O O B-geo O O O O O O O O B-gpe O O <- Predicted
O O O B-geo O O O O O O O B-geo O O O O O O O O B-gpe O O <- Actual
Accuracy = tensor(1.)

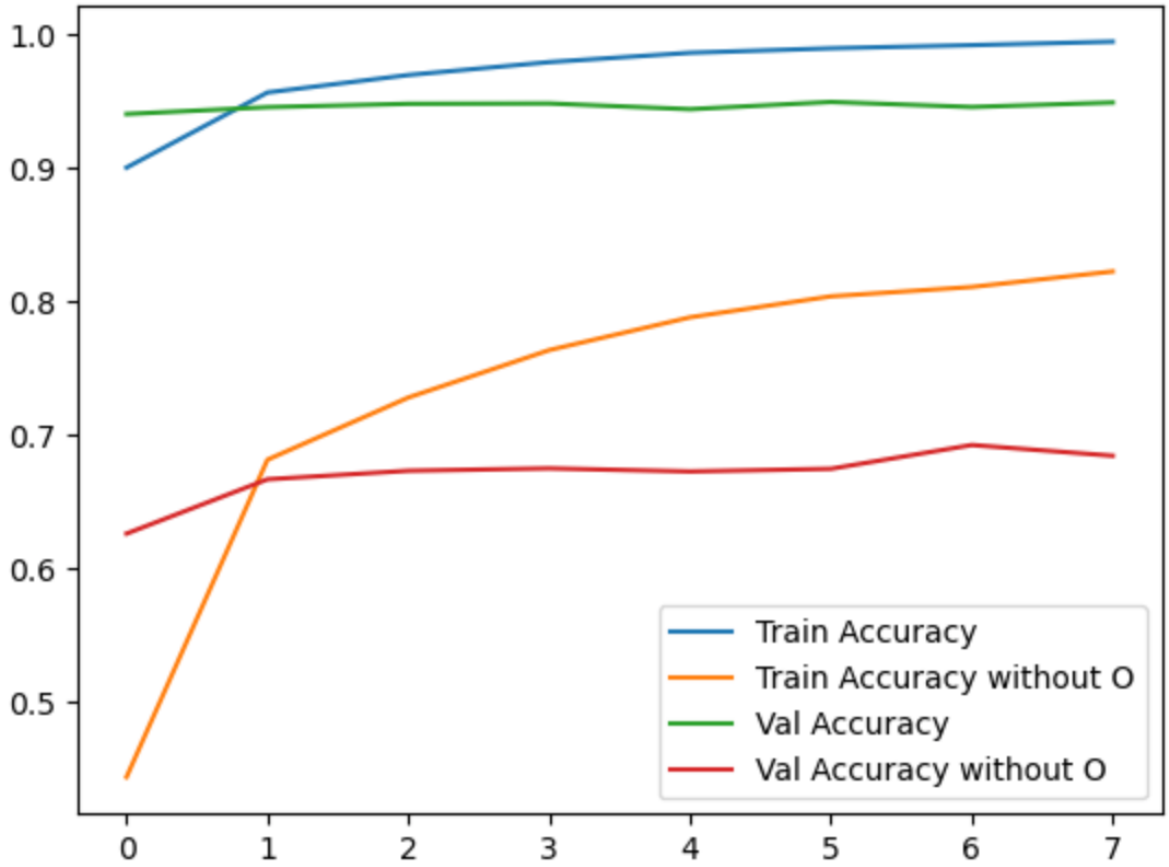
The blast shattered windows and caused the ceiling to collapse .
O O O O O O O O O O O <- Predicted
O O O O O O O O O O O <- Actual
Accuracy = tensor(nan)

Official ##s also blamed the weather for poor visibility that led to traffic accidents .
O O O O O O O O O O O O O O O <- Predicted
O O O O O O O O O O O O O O O <- Actual
Accuracy = tensor(nan)

The price of a barrel of oil for future delivery fell 35 cents [about one percent] to $ 3 8 . 68 a barrel during trading in New York .
O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O B-geo I-geo O <- Predic ted
O O O O O O O O O O O B-tim O O O O O O O O O O O O O O O O O O O O O O O O O O O B-geo I-geo O <- Ac tual
Accuracy = tensor(0.6667)

2) Using 5000 examples





Test Set Examples

Cambodian authorities have detained a fifth suspect in the 12 - year - old kidnapping and murder case of two men who were clearing land mines .
B-gpe O O O O O O O O O O O O O O O O O O O O O O O O O O <- Predicted
B-gpe O O O O O O O O O O O O O O O O O O O O O O O O O O <- Actual
Accuracy = tensor(1.)

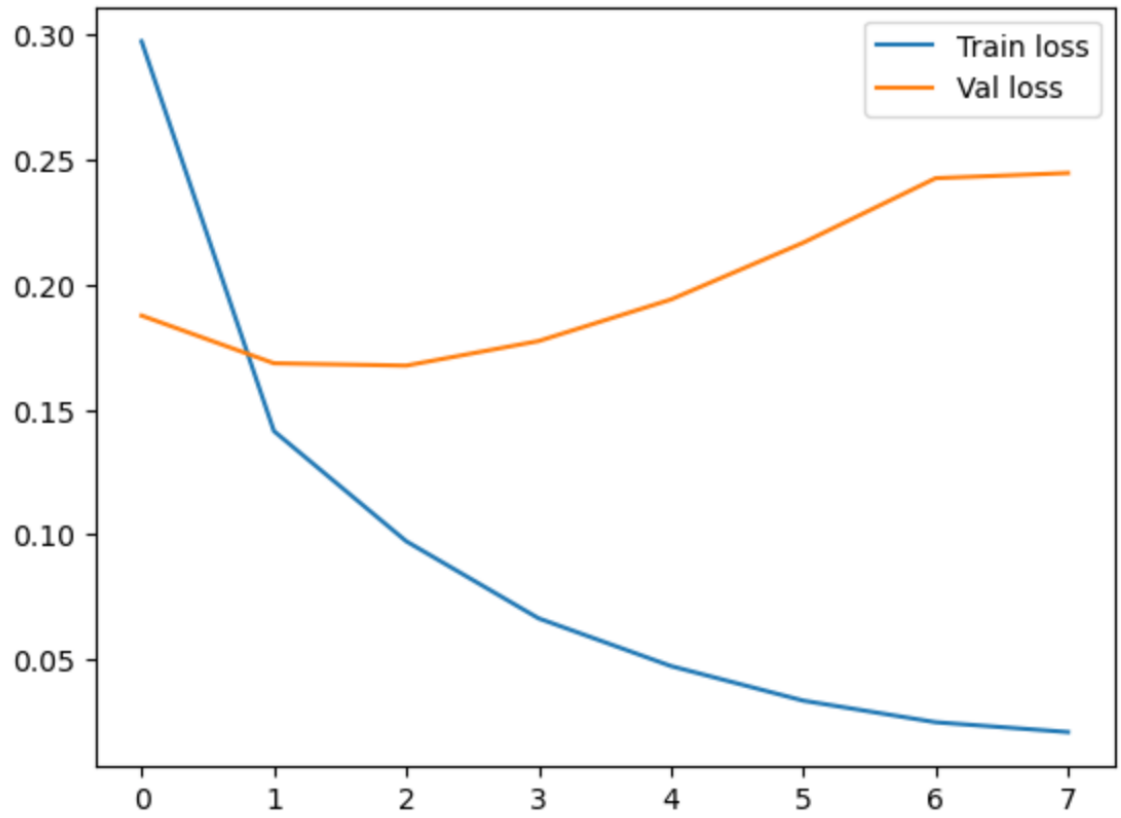
He ##lman ##d province has been a hot ##bed of ins ##urge ##nt activity since U . S . - led forces ou ##sted the hard - line Islam ##ist Tale ##ban rulers of Afghanistan in la te 2001 , following the September 11th terrorist attacks in the United States .
B-geo B-geo B-geo O O O O O O O O O O O O O O O O O O O O O O O O O B-org B-org I-org I-org O O B-geo O O B-tim O O O B-tim I-tim O O O O B-geo I-geo O <- Predicted
B-geo B-geo B-geo O O O O O O O O O O O O O O O O O O O O O O O O O B-org B-org I-org I-org O O B-geo O O B-tim O O O B-tim I-tim O O O O B-geo I-geo O <- Actual
Accuracy = tensor(1.)

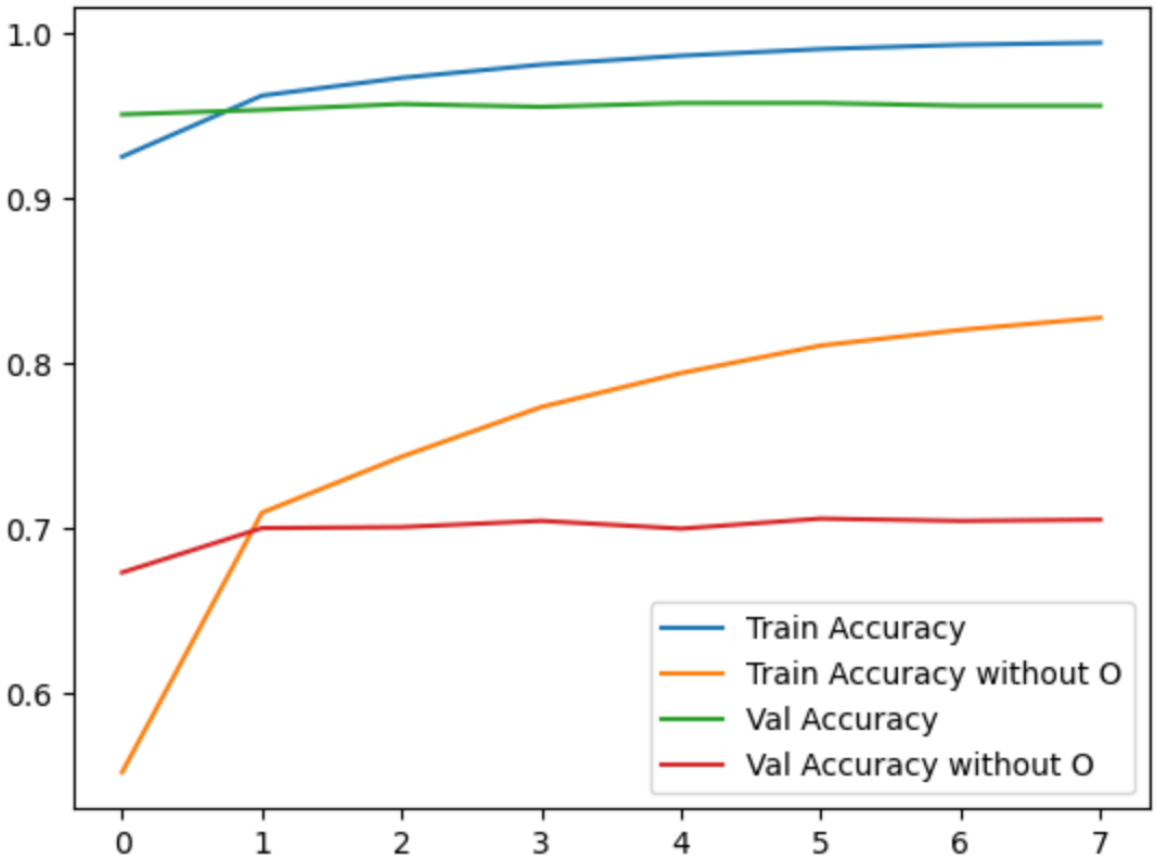
A ##vian flu has killed or forced the slaughter of millions of birds over the last two ye ars .
O O O O O O O O O O O O O O B-tim O O <- Predicted
O O O O O O O O O O O O O O B-tim O O <- Actual
Accuracy = tensor(1.)

Japan ' s Chief Cabinet Secretary Hi ##roy ##uki Ho ##so ##da says it is a necessary pr e ##con ##dition for any international match that players can play safely and spectat ors can be protected .
B-geo O O O O O B-per B-per I-per I-per I-per I-per O O O O O O O O O O O O O O O O O O O O O O O O O O O O O <- Predicted
B-geo O O O O B-per I-per I-per I-per I-per I-per I-per O O O O O O O O O O O O O O O O O O O O O O O O O O O O O <- Actual
Accuracy = tensor(0.6250)

An American general , Major General Joseph Peterson , told a U . S . daily , The Chica go Tribune , that the men were stopped by an Iraqi check ##point in northern Iraq las t month .
O B-gpe O O O B-per I-per I-per O O O B-geo B-geo B-geo B-geo O O B-org I-org I-org O O O O O O O O B-gpe O O O O B-geo O O O <- Predicted
O B-gpe O O O B-per I-per I-per O O O B-geo B-geo B-geo B-geo O O B-org I-org I-org O O O O O O O O B-gpe O O O O B-geo O O O <- Actual
Accuracy = tensor(1.)

3) Using 10000 examples





Test Set Examples

It was instituted in 1968 by the Bank of Sweden in memory of the founder of the Nobel P rize , Swedish philanthropist Alfred Nobel .
O O O O B-tim O O B-org I-org I-org O O O O O O O B-org I-org O B-gpe O B-per I-org O <- Predicted
O O O O B-tim O O B-org I-org I-org O O O O O O O B-org I-org O B-gpe O B-per I-per O <- Actual
Accuracy = tensor(0.8889)

Based on the WHO ' s records for Asia , Cambodia has the highest percentage of people i nfected with HIV - at 1 . 9 percent of the population .
O O O B-org O O O O B-geo O B-geo O O O O O O O O B-org O O O O O O O O O O <- Pre dicted
O O O B-org O O O O B-geo O B-geo O O O O O O O O B-org O O O O O O O O O O <- Act ual
Accuracy = tensor(1.)

Two other provisions in the bill would create a so - called guest worker program and wo uld de ##port illegal immigrants who have been in the country for fewer than five years .
O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O B-tim O O <- Predicted
O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O B-tim O O <- Actual
Accuracy = tensor(1.)

Apple ' s competitor , Microsoft , is developing a touch screen for the next version of its Windows software .
B-org O O O O B-org O O O O O O O O O O O B-org O O <- Predicted
B-org O O O O B-org O O O O O O O O O O O B-org O O <- Actual
Accuracy = tensor(1.)

The government has taken measures to curb violent crime , and recently adopted a fiscal reform package aimed at reducing the large gray economy and attracting foreign invest ment .
O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O <- Predicted
O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O <- Actual
Accuracy = tensor(nan)

Venezuela is a critical trade partner for the Caribbean island nation , which imports nearl y 1 , 00 , 000 barrels of oil per day from Venezuela .
B-geo O O O O O O O B-geo O O O O O O O O O O O O O O O O O O O O O O O B-geo O <- Predicted
B-geo O O O O O O O B-geo I-geo O O O O O O O O O O O O O O O O O O O O O O O B-geo O <- Actual
Accuracy = tensor(0.7500)

Conclusion

There's a good accuracy without taking O's which suggests that the model is learning the distribution of tags nicely.

The base BERT + Linear model could be improved by adding a CRF layer on top. I tries this but I was facing problems with the Kaggle GPU so I haven't attached any results.