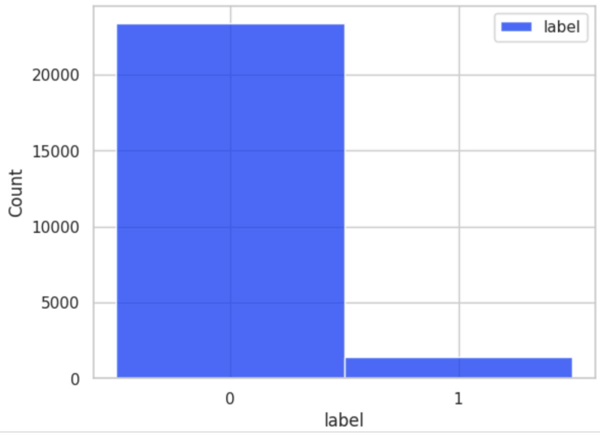
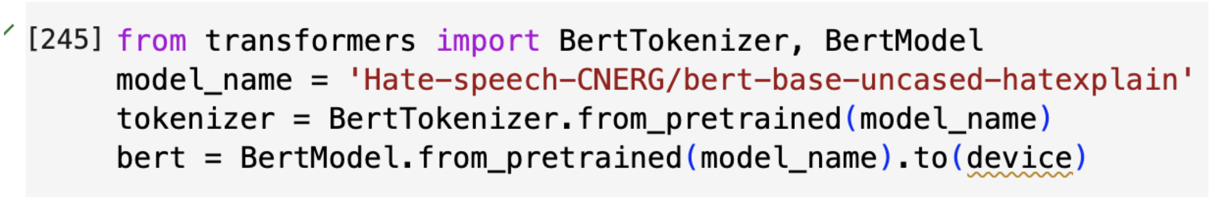
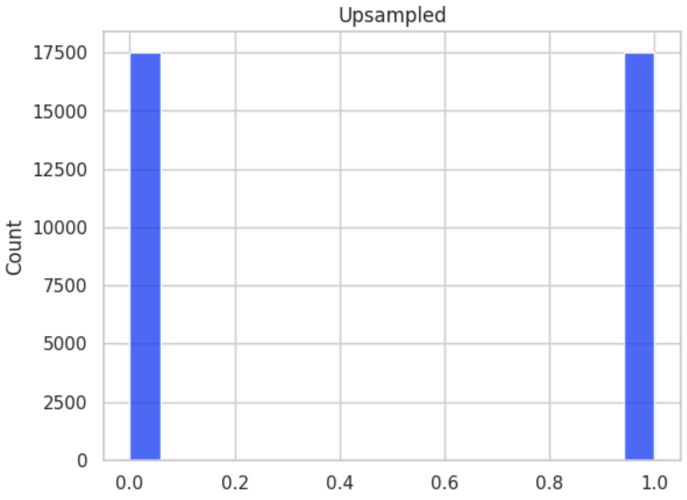


**We can see the hatespeech has very low percentage in the dataset, so we'll have to do upsampling. Also majority text has a length under 60. So we'll perform padding and truncation to unsure they all have a length 60, for tensor training by deep learning models.**

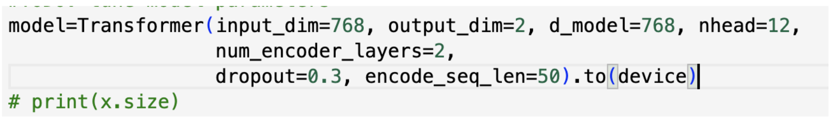
We use Bert embeddings from a pretrained model to generate tokens and embeddings on our dataset.



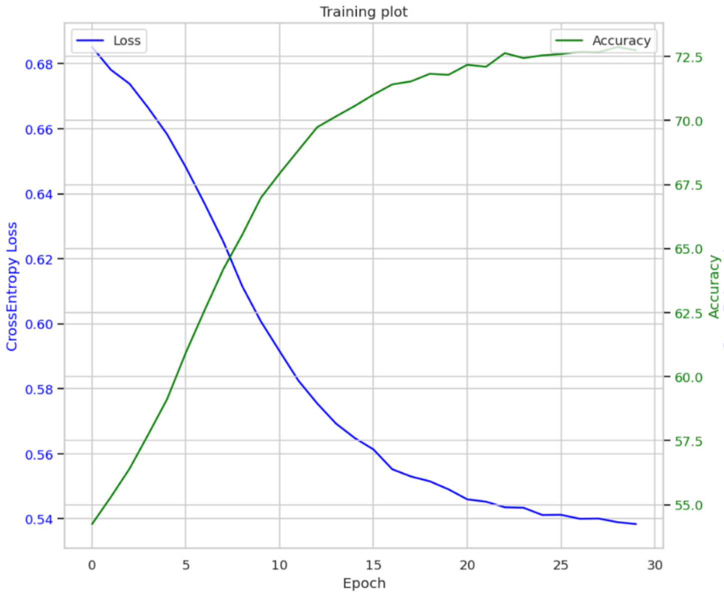
```
[245] from transformers import BertTokenizer, BertModel
      model_name = 'Hate-speech-CNERG/bert-base-uncased-hatexplain'
      tokenizer = BertTokenizer.from_pretrained(model_name)
      bert = BertModel.from_pretrained(model_name).to(device)
```

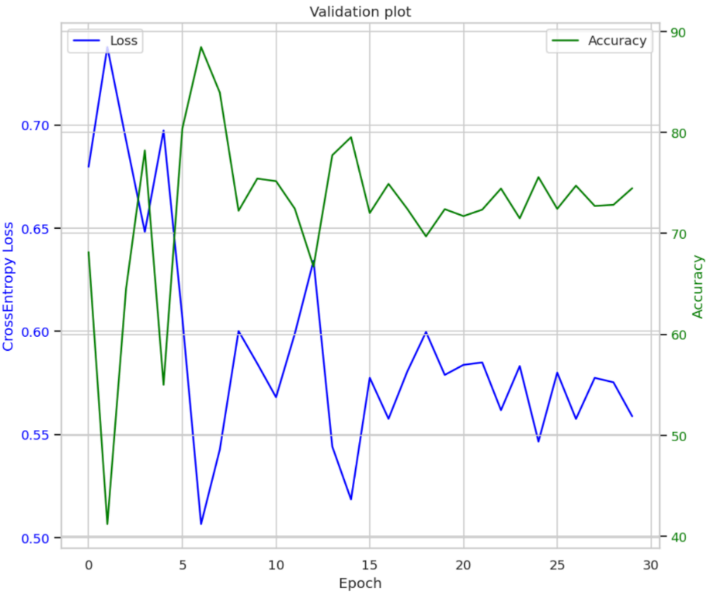


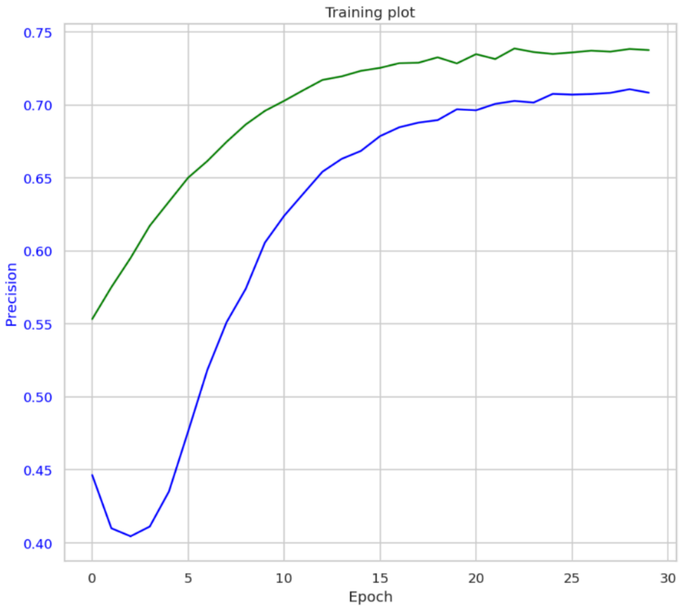
We use Transformer Encoder model to train

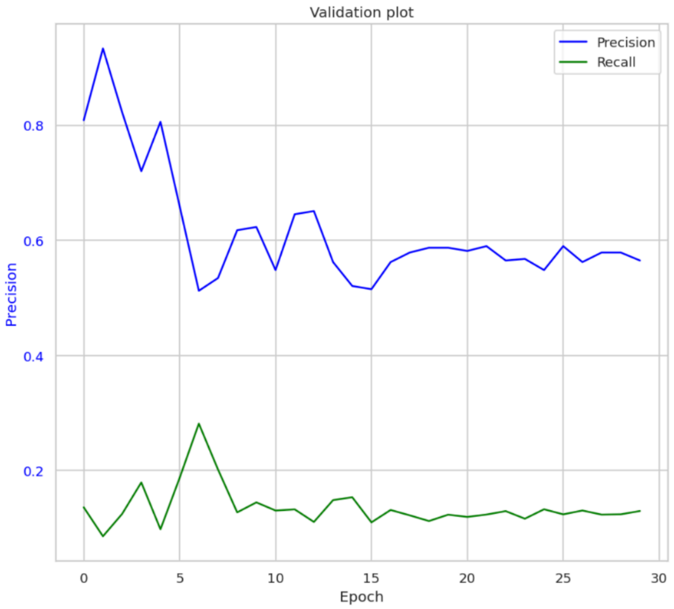


```
model=Transformer(input_dim=768, output_dim=2, d_model=768, nhead=12,
                  num_encoder_layers=2,
                  dropout=0.3, encode_seq_len=50).to(device)
# print(x.size)
```







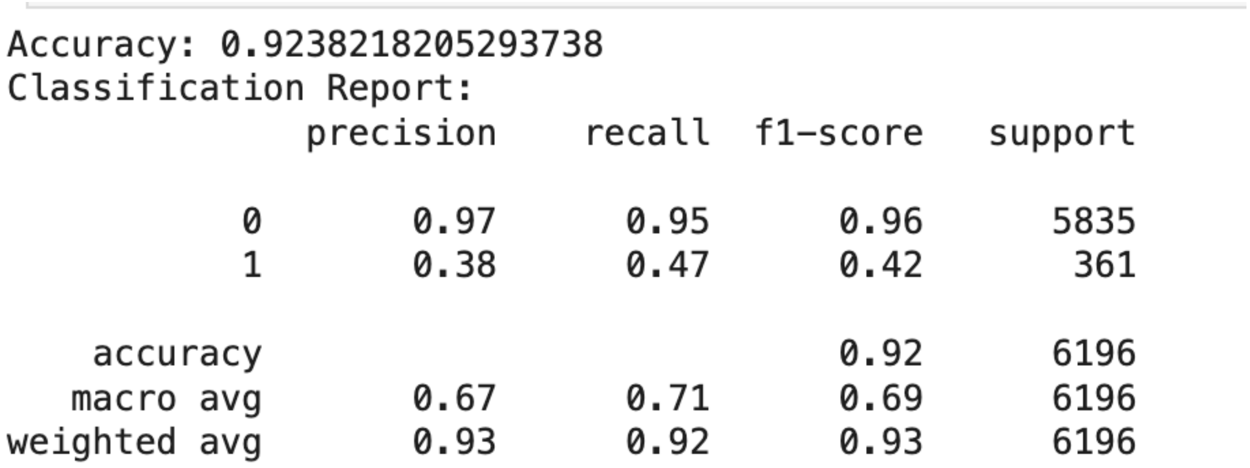


**Because Hate Speech consists only 5% of our dataset. Predicting all data as non-hate can give you a accuracy 95%, but we fail to detect hateful speech. Therefore, we have to balance precision and recall. The model at the beginning has a high precision around 90% percent by predicting almost all data to be 0(non-hate), and the recall is around 5%(fail to detect hate). By training the model, accuracy is getting lower then rise and we are getting a lower precision higher accuracy and recall**

**Training: Precision 0.7074 Recall 0.7372**
**Validation: Precision 0.5623 Recall 0.1306**

The recall does not increase a lot says that our model does not increase its ability in detecting hate speech very much.

***Random Forest with Bert Embeddings***

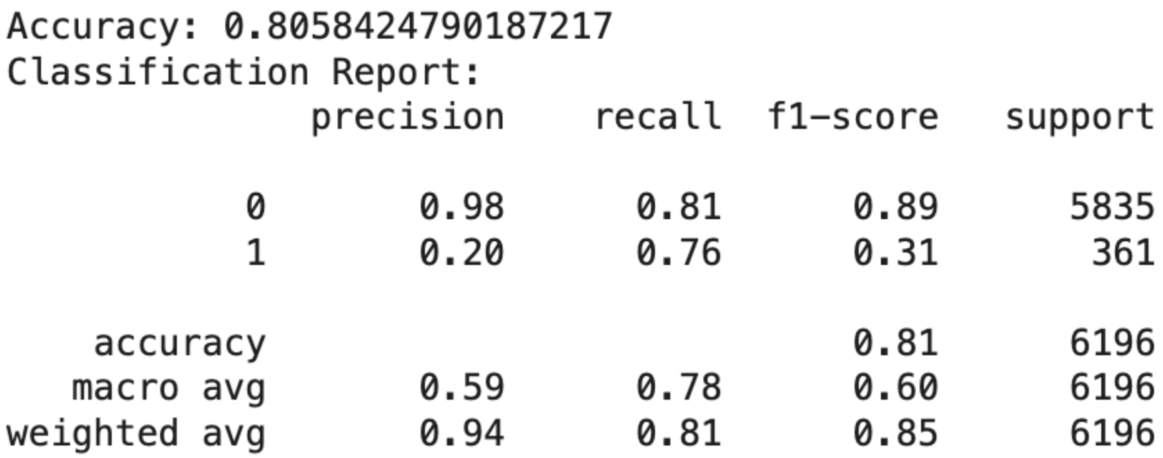


```
Accuracy: 0.9238218205293738
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.95      0.96      5835
           1       0.38      0.47      0.42       361

    accuracy                           0.92      6196
   macro avg       0.67      0.71      0.69      6196
weighted avg       0.93      0.92      0.93      6196
```

***XGBoost with Bert Embeddings***



```
Accuracy: 0.8058424790187217
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.81      0.89      5835
           1       0.20      0.76      0.31       361

    accuracy                           0.81      6196
   macro avg       0.59      0.78      0.60      6196
weighted avg       0.94      0.81      0.85      6196
```

**The results of Random Forest(0.38, 0.47) XGBoost(0.2, 0.76) for (precision, recall) suggests that tree-based models on our task has a lower precision and a higher recall that Neural Network I trained. Tree-based models with Bert embeddings pretrained have a ourstanding ability to detect hateful speech suggests that bert embeddings is powerful on downtream tasks.**