Named Entity Recognition
Species - 800

- Constantin Gabriel-Adrian
- Sociu Daniel
- Onescu Iancu-Gabriel

Task description & Dataset exploration

- The chosen task was Named Entity Recognition with the aim of recognizing tokens that mention organisms in texts that belong to eight different categories:

  - ❏ bacteriology
  - ❏ botany
  - ❏ entomology
  - ❏ medicine
  - ❏ mycology
  - ❏ protistology
  - ❏ virology
  - ❏ zoology

Dataset	None cls (0)	Species (1)	Species+ (2)
Train	141424	2557	3310
Valid	21348	384	485
Test	40488	767	1043

Main approach

- Considering the fact that we have to solve a token classification task into categories we had a few main steps to follow:
  1) We obtain the embeddings for each token to later on use a classifier to properly classify them.
  2) Train a transformer model for the classification of our embedded tokens
  3) Predict on the test data
- Therefore in the following slides we will present some implementation details and the obtained results.

BERT

- To train the classifier, we must process the tokens such that the data will be represented in numerical form. Thus, we used BERT transformer to transform the tokens because:

1) It takes into account the context of the sentence
2) It's pretrained on a dataset of considerable size
3) Has variations that were pretrained on biology-specific data
4) Has variations that were pretrained on NER-specific tasks

Tokenization & padding

- Considering we have sentences of different lengths (number of tokens), and that we are using a transformer, we have to make each sentence a specific length.
- We decided to have each sentence with 128 tokens, therefore we either split it (if it is longer) or we pad it.
- To be mentioned that the tokenizer used splits some tokens with multiple sub-tokens, which should not be learned by the transformer, therefore we assigned them a special label (-100).

Training via Feature extraction

- Given the fact that the aforementioned model was trained on an entirely different task & dataset, we had to find a way to adapt it to our challenge.
- Thus, in the literature there are 2 main approaches:
  1) Continue model training using the pretrained weights as the starting point
  2) Feature Extraction – freeze embedding layers in order to preserve model's capacity to represent feature's data in latent space
- We explored the option of retraining some layers (mainly classifier layers) of the model in order to fine-tune our solution.

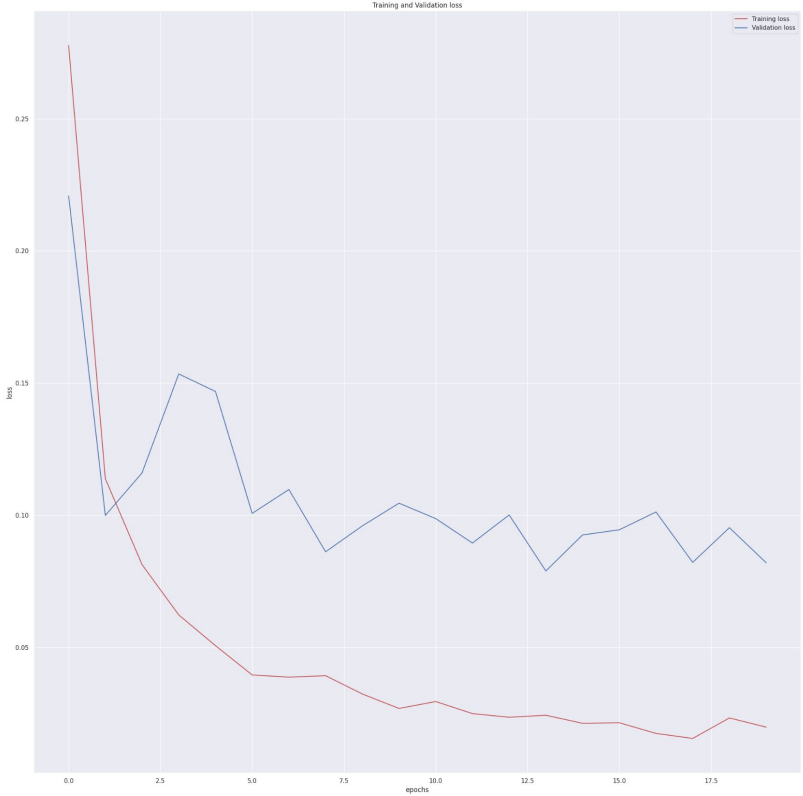
Final results - all our attempts

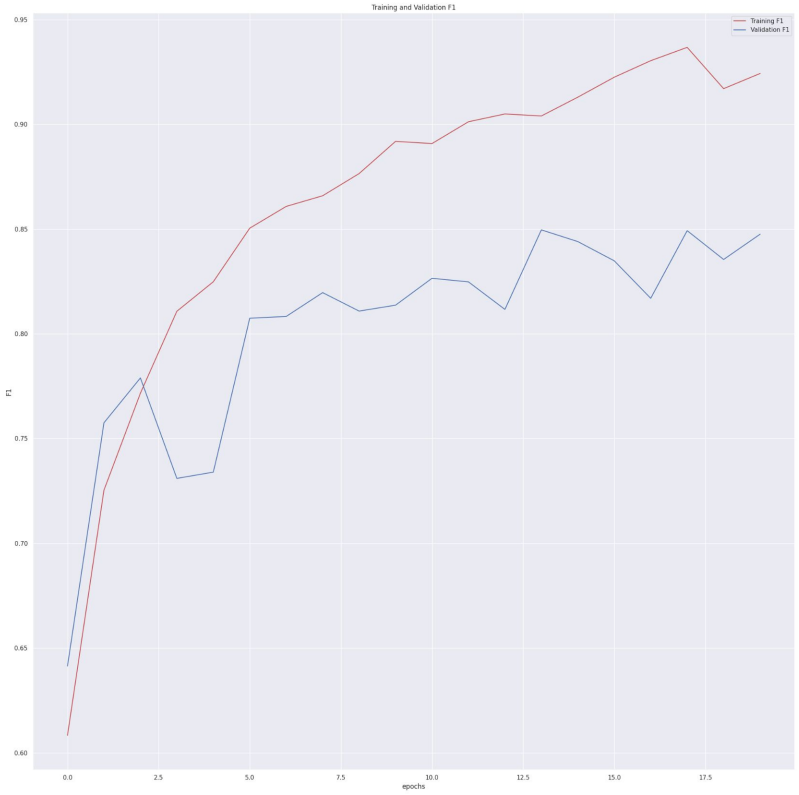
Model	Loss	Accuracy	F1 score
BERT uncased	0.105	0.953	0.816
RoBERTa	0.123	0.938	0.817
BertNER	0.119	0.939	0.828
BERT cased	0.086	0.944	0.850
BioBERT cased	0.089	0.944	**0.866**

Final results - our model vs literature

Model	Loss	Accuracy	F1 score
BioKMNER + BioBERT	-	-	0.763
SciFive-Base	-	-	0.765
BioFLAIR	-	-	0.824
Spark NLP	-	-	0.825
BioBERT cased	0.089	0.944	**0.866**

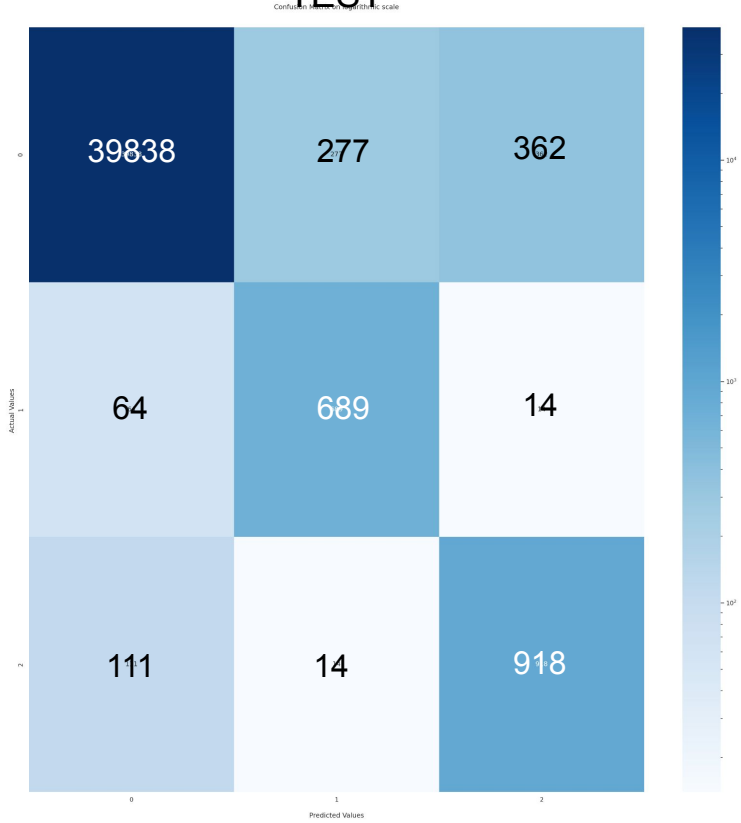
Loss and f1 values for the best model



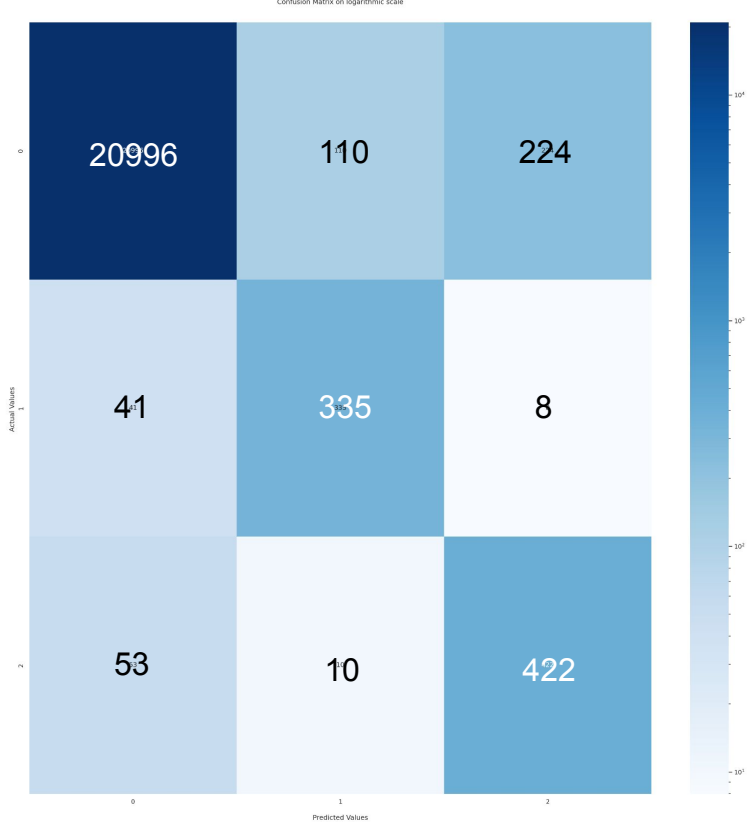


Confusion matrix for the best model

TEST



VALID



Thank you for your **attention**!