Project 2

Gabriel Ferreira
gabferre@iastate.edu
GabrielGomesFerreira

1 Methods

This section details the approach employed for developing and evaluating a binary document classification using BERT model. The task involves automatically identifying research documents relevant to animal QTL. The methodology described here includes dataset, data pre-processing, BERT classification model, training configuration and procedures, and evaluation metrics.

1.1 Dataset

To facilitate this binary classification task, two datasets were provided: (1) a labeled dataset used for model training phase, and (2) an unlabeled dataset intended for evaluation. We define the datasets as:

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where each x_i denotes a research document, and each label $y_i \in \{0, 1\}$ indicates whether the document is related (1) or unrelated (0) to animal QTL.

The labeled dataset for training contains $N = 11,278$ labeled observations, each containing four columns: 'PMID', 'Title', 'Abstract', and 'Category'.

Similarly, the unlabeled dataset used for evaluation contains $M = 1,097$ observations with an identical structure, except the relevance field is labeled as 'Label' and initialized with placeholder zeros (0) pending the model's predictions.

1.2 Pre-Processing for BERT

Since BERT requires specific input formatting, the following pre-processing steps were performed to prepare the textual data for classification using BERT:

- **Text Concatenation:** The title and abstract fields were concatenated with a blank space (" ") separator into a single input text for each document, providing comprehensive context to the model.
- **Tokenization with WordPiece:** Each input text was tokenized using BERT's WordPiece tokenizer, converting raw text into subword tokens compatible with BERT's vocabulary.
- **Padding and Truncation:** Token sequences were truncated or padded to a uniform maximum sequence length (512), ensuring consistency across model inputs.
- **Attention Masks:** Attention masks were created to differentiate meaningful tokens from padded tokens, allowing BERT to ignore padded elements during training.
- **Train-Validation-Test Splitting:** The dataset is split into training (80%) and validation (20%) sets. The splits were performed in a stratified manner to maintain consistent class proportions across each subset, thereby ensuring reliable model training and validation.

1.3 Classification Model (BERT-based)

The best-performing classification model in our experiments was Bidirectional Encoder Representations from Transformers (BERT), a transformer-based neural network architecture known for effectively capturing deep contextual representations from text.

Unlike traditional sequential models, BERT processes text bidirectionally, simultaneously considering both preceding and succeeding contexts. This characteristic enables it to grasp nuanced relationships within language data.

Specifically, we utilized the BERT-base-uncased variant, which comes pre-trained on a large corpus of English text and then we subsequently fine-tuned it using the training dataset specific to our classification task.

1.4 Training Parameters and Configuration

The fine-tuning of our BERT model involved several parameters and configurations to achieve better performance. Specifically, the model was trained using a weighted loss function to address class imbalance observed in the training data. Class weights were computed using sklearn's *compute_class_weight* with a balanced scheme to ensure equal representation across classes during training.

The following hyper-parameters and settings were employed during training:

- Learning Rate: $\{2 \times 10^{-7} \leq x \leq 2 \times 10^{-4}\}$
- Batch Size: $\{16, 30\}$ samples per device
- Number of Epochs: 15
- Optimizer: AdamW (default optimizer in Hugging Face Transformers)
- Loss Function: Weighted Cross-Entropy Loss, incorporating computed class weights
- Metrics evaluated during training and validation included: F1-score, Precision, Recall

Model checkpoints were saved at each epoch, and the best-performing model (based on validation set performance) was retained. The threshold for converting model output probabilities into binary predictions was set at 0.4 to optimize classification performance metrics, particularly recall and F1-score.

All training and evaluation routines were implemented using Hugging Face's Trainer API, extended to support weighted loss computation via a custom subclass.

The training execution was conducted on the Kaggle Notebook environment, utilizing GPU acceleration for efficient computation.

2 Results

The fine-tuned BERT model was evaluated using both an internal evaluation set and an external test set from a Kaggle competition. The primary evaluation metric employed was the F1-score, calculated as the harmonic mean of precision and recall.

The results from the internal evaluation and Kaggle Leaderboard are described next.

2.1 Validation Set (internal test set):

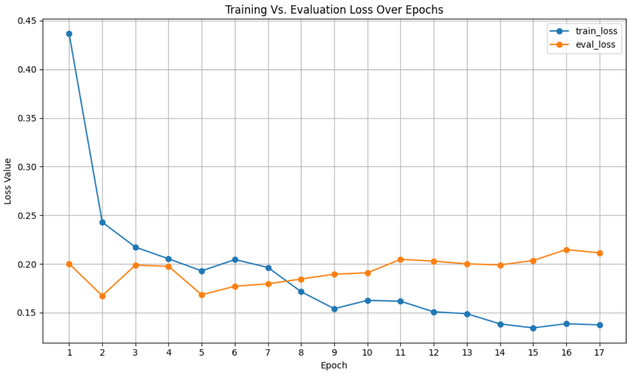

Figure 1: Train Vs. Evaluation Loss

The training loss decreased steadily and settled at a low value, indicating that the model was effectively learning from the data. The validation loss stabilized early and remained consistently low, suggesting good generalization. After epoch 10, a slight divergence between training and validation losses emerged, indicating early signs of overfitting to the training set, which guided our decision to stop training at that stage.

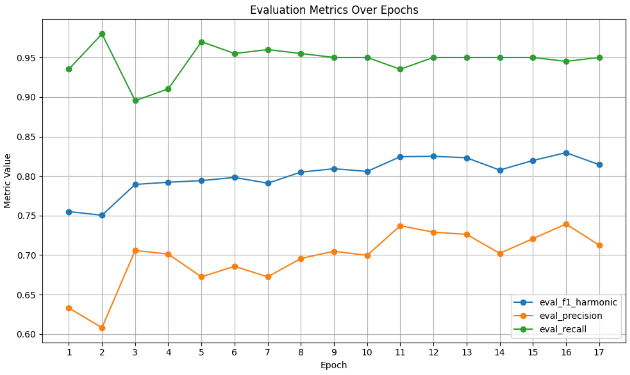

Figure 2: Evaluation Chart

The classifier demonstrated steady improvements across epochs, achieving an F1 score stabilizing around 80%, indicating a good balance between precision and recall. Precision showed a noticeable upward trend from approximately 62% to near 75%, while recall consistently remained high, around 95%. These results highlight the model's robust capability to identify relevant papers, which aligns well with the goals of this task—supporting curators by prioritizing recall, as including an irrelevant paper (false positive) is more acceptable than missing a relevant one (false negative).

2.2 Kaggle Competition (external test set)

The performance of the final model was evaluated externally through a Kaggle competition. On the Public Leaderboard, which comprised 29.9% of the total test data, the model achieved an F1-score of 84%. This result represents a substantial improvement of 14% compared to the baseline model provided by the professor, which achieved an F1-score of 70%. This result indicates strong overall performance and confirms the model's effectiveness in generalizing to unseen data.

References

[1] Hugging Face. Transformers Library. Available at: https://huggingface.co/transformers/.

[2] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://arxiv.org/abs/1810.04805.

[3] Venelin Valkov. Fine-Tuning BERT for Text Classification with Hugging Face. Available at: https://www.youtube.com/watch?v=qWfzIYCvBqo.

[4] CodeEmporium. What is BERT? | BERT Explained. Available at: https://www.youtube.com/watch?v=IzbjGaYQB-U.

[5] Data School. How to Handle Imbalanced Classes in Machine Learning. Available at: https://www.youtube.com/watch?v=4QHg8Ix8WWQ.

[6] Abhishek Thakur. Using Class Weights for Imbalanced Dataset. Available at: https://www.youtube.com/watch?v=9he4XKqqzvE.

Assistance from AI (ChatGPT) was used for refining the structure and wording of this report.