Conspiracy detection PAN 2024

Bobic Teona-Christiana and Dobrin Ionut

University of Bucharest, Romania
Bitdefender, Romania

bobic.teona20@gmail.com, ionutdobrin2003@gmail.com, fbrad@bitdefender.com, ipintilie@bitdefender.com

1. Introduction

Given a text from the PAN24 Oppositional Thinking Analysis dataset, classify it as either CONSPIRACY (misinformation, labeled FALSE) or CRITICAL (credible, labeled TRUE). Challenges of this task include:

1. Fine-tuning a multilingual BERT model
2. Compare Multilingual BERT to English Bert
3. Perform feature analysis

2. Dataset and Preprocessing

Dataset Description:

The dataset used in this project is sourced from the PAN24 Oppositional Thinking Analysis, focusing on conspiracy theories vs critical thinking narratives. It consists of three subsets: training, validation, and test sets. Each article includes a id, text body, and a category label (TRUE or FALSE), which are mapped to CRITICAL (credible, 1) and CONSPIRACY (misinformation, 0).

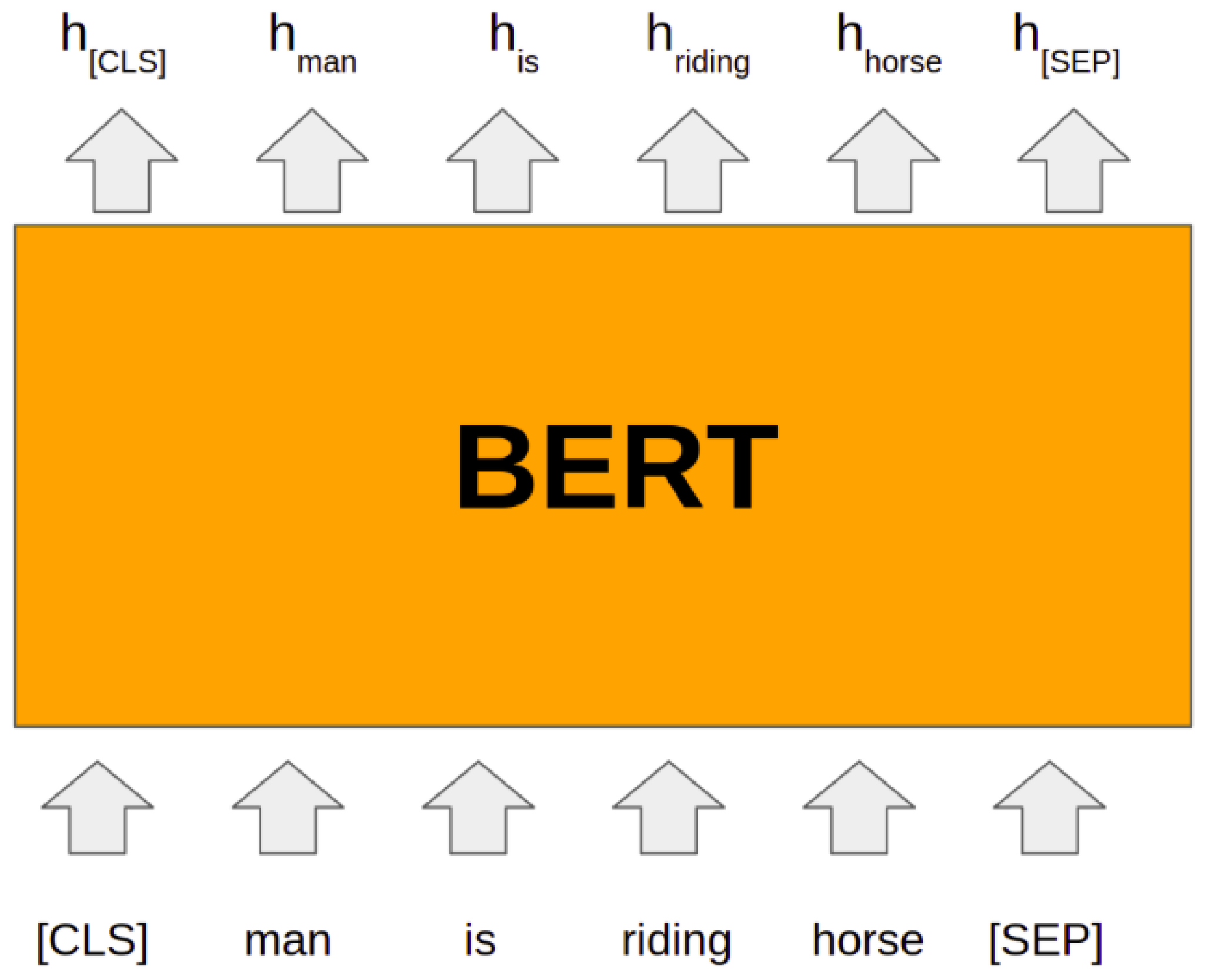
The articles vary in length, with approximately 614 characters per text.

Data preprocessing:

The text was cleaned by removing non-ASCII characters, URLs, and excessive whitespaces. Labels were mapped from categories (CONSPIRACY to 0, CRITICAL to 1), and rows with missing labels were removed, resulting in 4000 valid samples (2621 CRITICAL, 1379 CONSPIRACY).

The dataset was split into training (70%, 2800 samples), validation (15%, 600 samples), and test (15%, 600 samples) sets. The training set, initially imbalanced (1835 CRITICAL, 965 CONSPIRACY), was balanced by oversampling the minority class (CONSPIRACY) to match the majority, yielding 3670 samples (1835 per class).

Texts were tokenized using two BERT tokenizers: bert-base-multilingual-cased and bert-base-cased. Tokenization included padding to a fixed length, truncation of long texts, and a maximum length of 512 tokens, incorporating special tokens ([CLS], [SEP]).



3. Models

Baseline:

We fine-tuned a pre-trained Multilingual BERT model (bert-base-multilingual-cased). It was adapted for sequence classification with two output labels: CONSPIRACY and CRITICAL.

In addition to this, we fine-tuned an English BERT model (bert-base-cased), which is specifically pre-trained on English text.

Hyperparameter Configuration:

1. Number of training epochs: 5
2. Batch size: 16 (for both training and evaluation)
3. Learning rate: 2e-5
4. Warmup steps: 100
5. Weight decay: 0.01
6. Dropout probabilities: 0.4 (for both hidden layers and attention mechanisms)

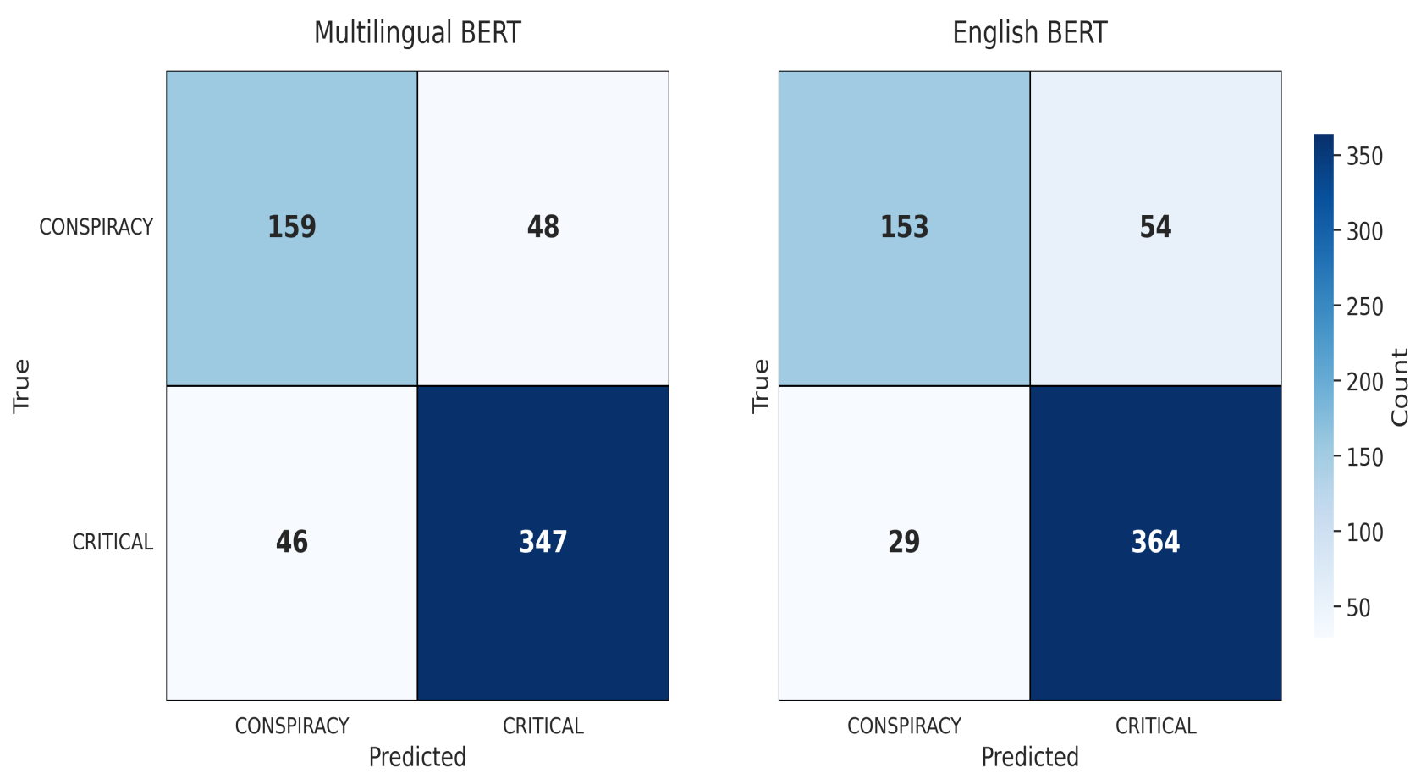
The fine-tuning process involved adjusting the pre-trained weights to better suit the specific task of distinguishing between CONSPIRACY and CRITICAL narratives.

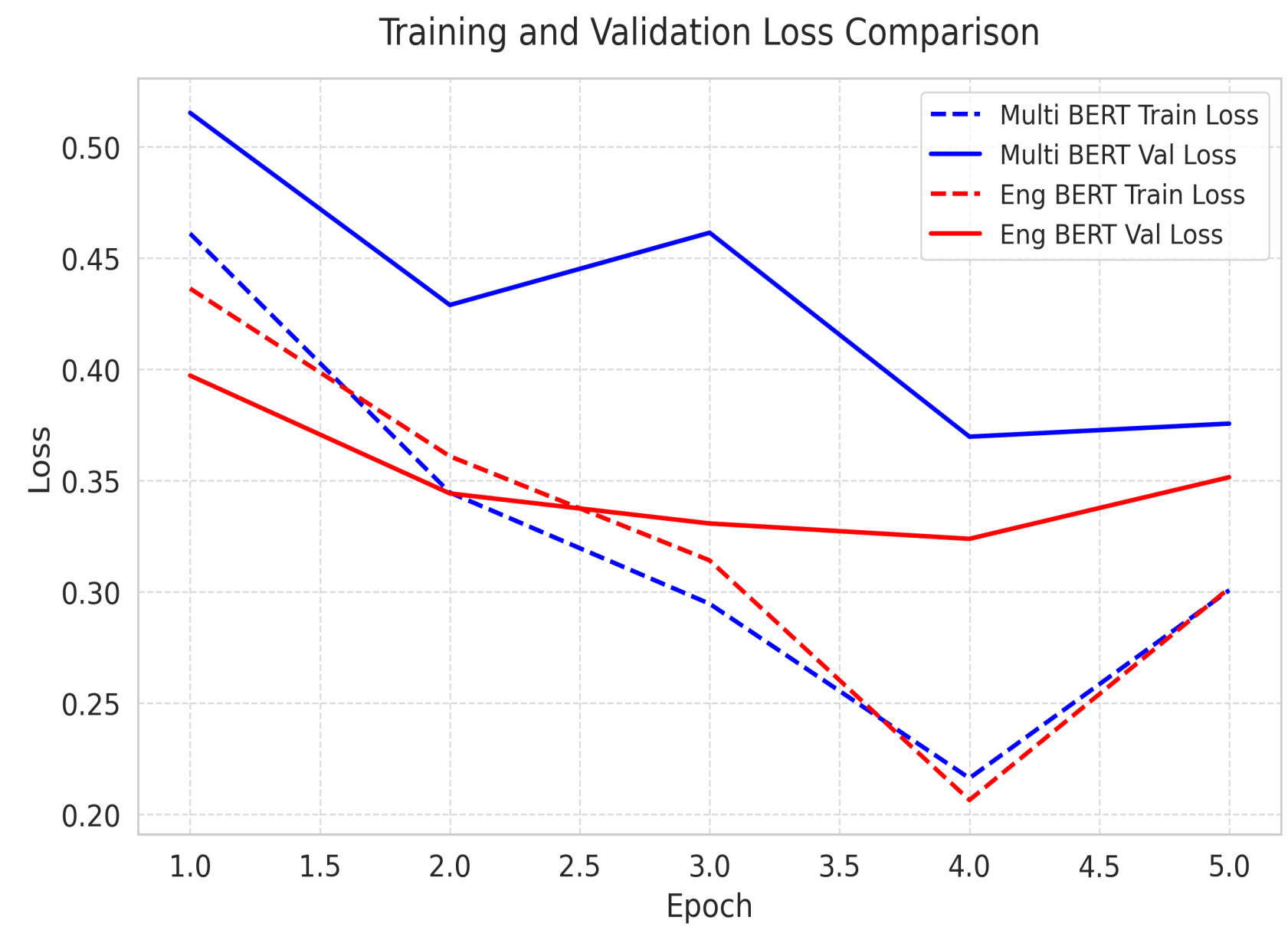
We implemented a custom WeightedTrainer to incorporate class weights, which were computed based on the training set's label distribution.

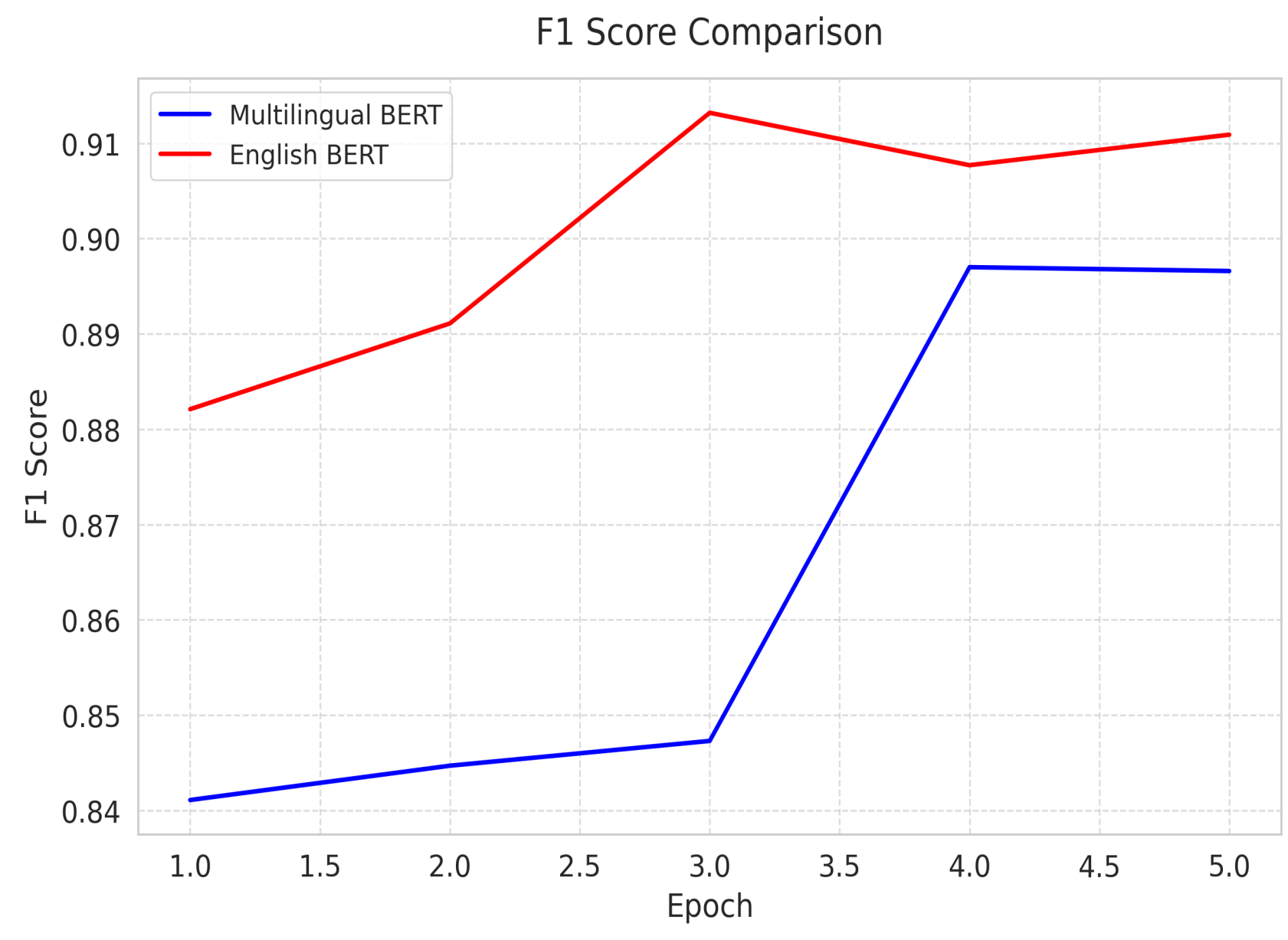
4. Results and Conclusion

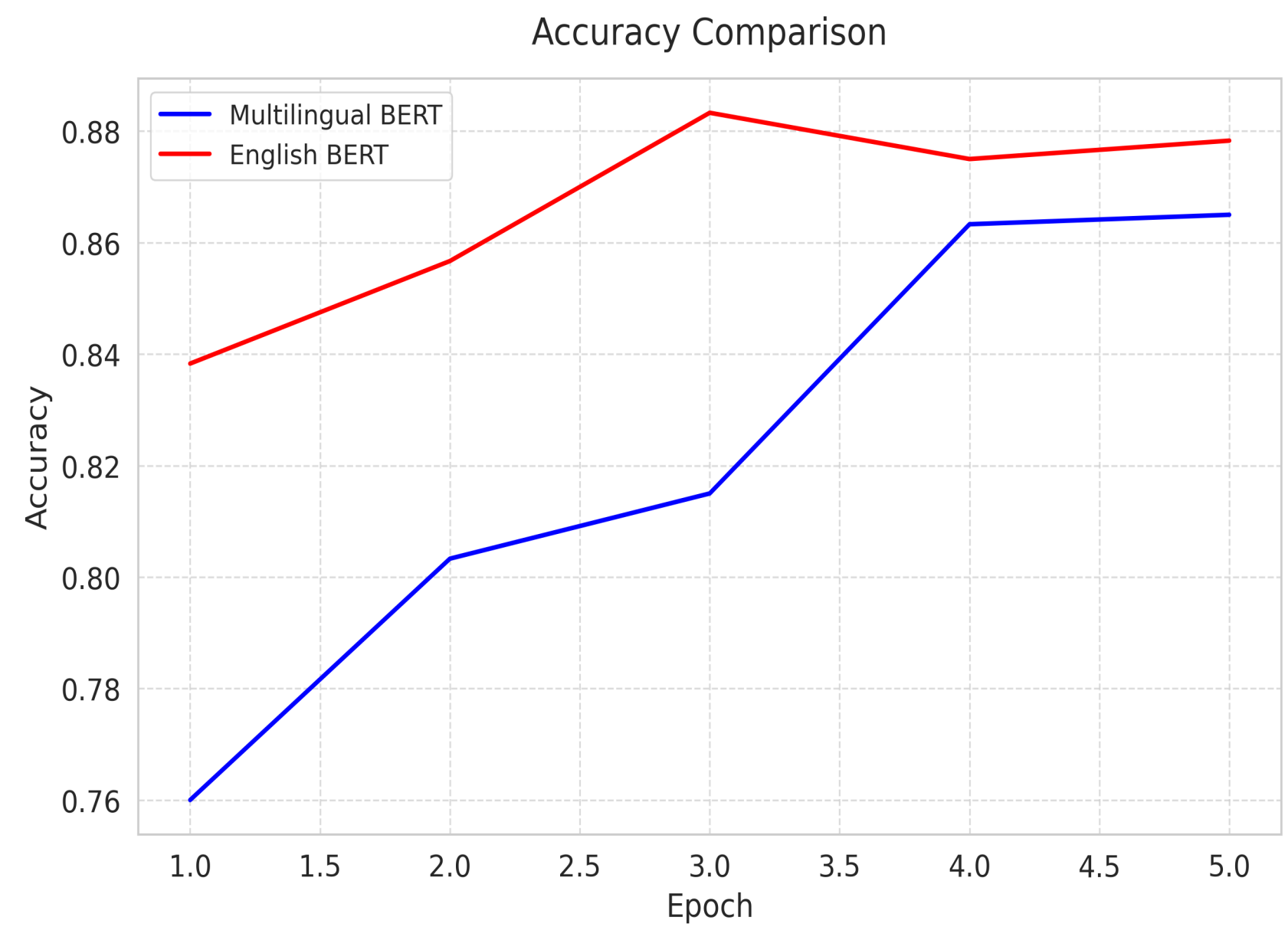
A comparison of the best-performing epochs:

- Multilingual BERT: MCC: 0.7024, Accuracy: 0.8650, F1: 0.8966
- English BERT: MCC: 0.7379, Accuracy: 0.8833, F1: 0.9132









Multilingual BERT FP Text

' My son died after getting the vaccine , but still get your kids vaccinated '

Multilingual BERT FN Text

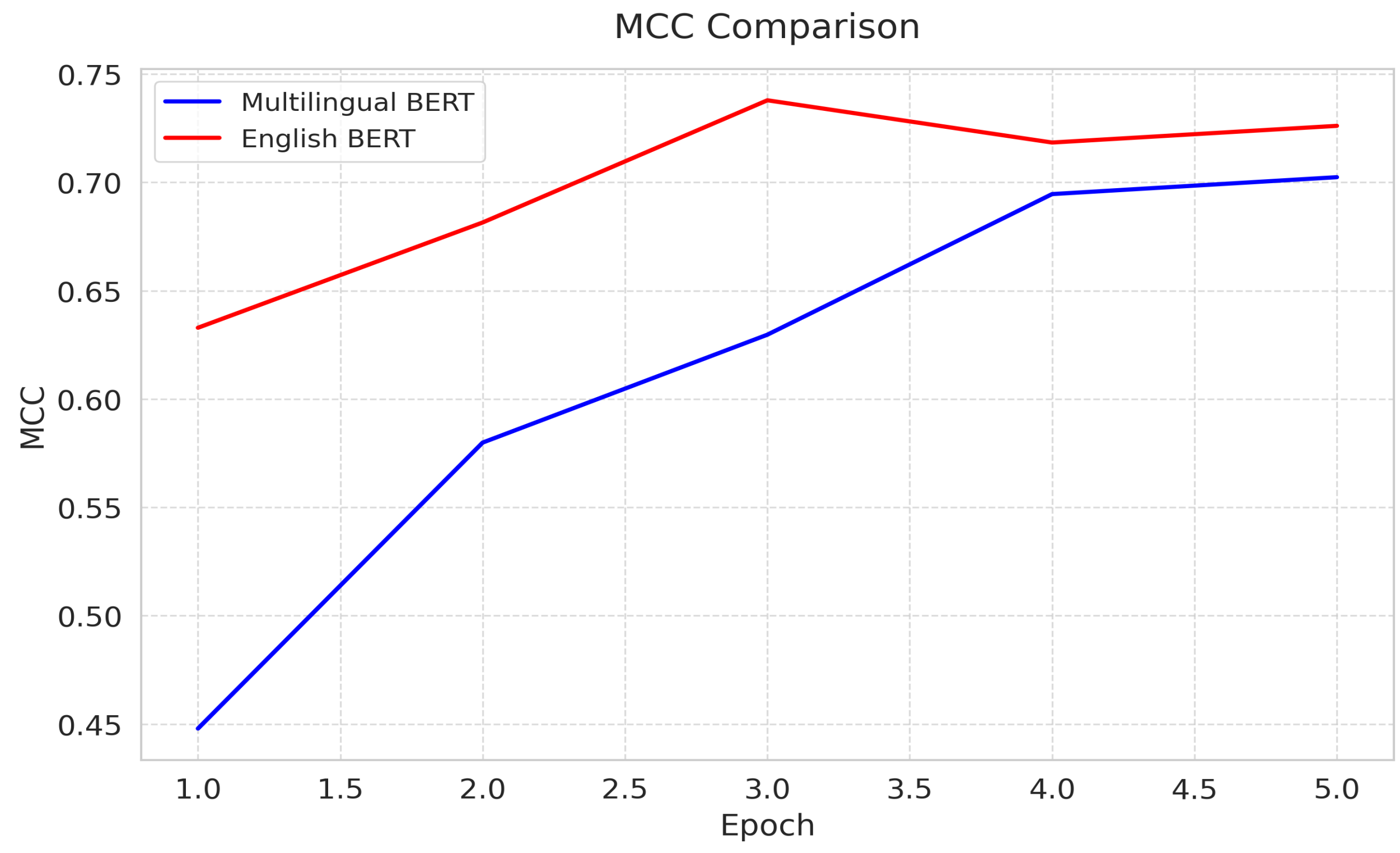
' US Military personnel spotted in the South Korean mass death event , wtf are they doing there ? Probably Checking the aftermath of the HAARP activation '

English BERT FP Text

'Dr. Roger Hodkinson is a decorated Canadian physician who stepped forward to declare that the COVID - 19 pandemic is more or less a hoax . He runs a biomedical testing company that sells COVID - 19 tests '

English BERT FN Text

' Still can't believe this is reality . It 's completely mind blowing at this point . STAY AWAY FROM THE VACCINE . We can not lay any more emphasis on this . I am pretty sure most of you have seen already '



Conclusion

The English BERT consistently outperformed the Multilingual BERT, likely due to its focus on English-specific linguistic features, making it the best model based on MCC. These findings demonstrate the English BERT's effectiveness while underscoring the need for bias mitigation to further refine model performance.