BERT FINETUNING FOR ARGUMENT CLASSIFICATION IN A GOVERNMENT CONTEXT

Project Summary 2025

METHOD

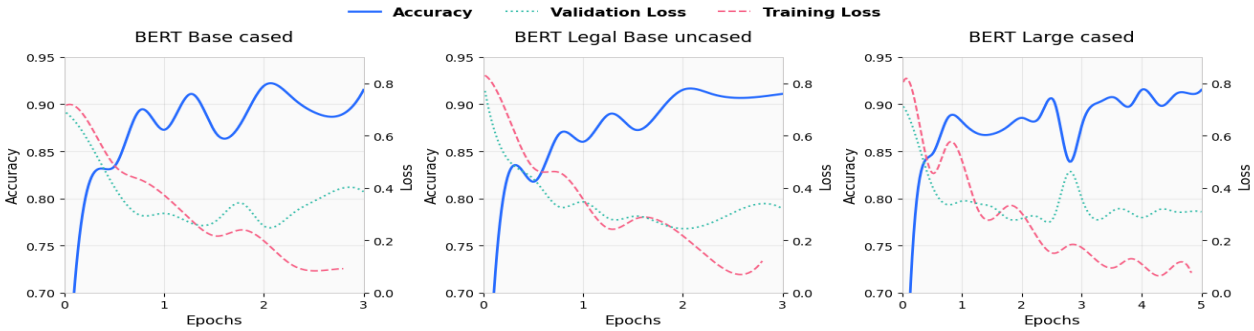
BERT is a deep, multi-layer language model based on a bidirectional transformer architecture and pre-trained to generate contextual language representations. BERT models are effective for text classification as their bidirectional design captures patterns and context from both preceding and following tokens, ensuring enhanced semantic understanding.

APPROACH

Three pretrained BERT models were identified as most relevant for experimentation: BERT Base model (cased), BERT Large model (cased) and BERT Legal Base model (uncased). A linear layer was added to the network as a classification head with output size 2, corresponding to the two target classes of arguments. A baseline for classification accuracy was established with the Bag-of-Words model representing the text as fixed-length vectors encoding the word frequency.

TRAINING PERFORMANCE

The best hyperparameters were identified using an automated optimization procedure. The figure below illustrates the evaluation accuracy along with the training and validation losses for the three BERT models finetuned with these optimal hyperparameters.



RESULTS

All three finetuned BERT models achieved a substantial improvement in classification accuracy (91-92%) compared to the Bag-of-Words baseline. The Base model slightly outperformed the Large and Legal Base variants, which was likely the result of randomness inherent in the hyperparameter search process. The achieved accuracy is summarized below:

Model	Type	Best Accuracy, %
Benchmark	Bag-of-Words	85.17
Base cased	BERT pretrained	91.95
Large cased	BERT pretrained	91.53
Legal Base uncased	BERT pretrained	91.53

The results confirm argument classification can be implemented based on linguistic properties of the text using modern computational approaches, such as bidirectional transformer-based language models. At the same time, the small size of the dataset (1,178 examples) raises concerns regarding the model's ability to generalize. The observed performance may not reliably translate to unseen examples.

CONCLUSION

The project demonstrates the finetuned BERT models are capable of performing well in tasks involving the detection of fallacious reasoning in government discourse. Ensuring the model can generalize well and is suitable for application in the real-life settings will require further development, including expanding the dataset, applying data augmentation, exploring more domain-relevant pretraining and experimenting with alternative architectures, embeddings and training strategies.

GitHub repository: https://github.com/EvgenyP1x/Argument_Classification_with_BERT/