PURDUE UNIVERSITY | The Data Mine

Hyper-Parameter Optimization for Natural Language Processing (NLP) Models

BATTELLE

INTRODUCTION

Battelle: Battelle is a private non-profit research organization that advances science and technology to have the greatest impact on our society and economy.

About ***Sematrix:*** It can take up to 17 years to translate medical studies into practice. The Sematrix NLP model will be used in Medical Literature and will save researchers much time reading through entire papers.

What is ***Natural Language Processing (NLP):*** A field of artificial intelligence that involves teaching machines to understand human language.

What is ***BERT:*** A powerful language model called to perform specific tasks like information extraction within NLP.

Our ***Goal:*** To extract crucial information from research and scholarly papers using the BERT transformer model.

Our ***Challenge:*** There are enormous amounts of medical text documents and data available. Their classification and the appropriate manifestation of their relations with each other become a necessary and tedious task. That is why we aim to fine-tune a BioBERT pre-trained language model, that has been pre-trained over large corpora of PubMed and PMC texts.

OUR MODELING PROCESS

First…

We used brat-parser to import the files and extract the named entity data to make the raw ANN files from Harvard readable by our model. Once the data was extracted, we wrote a custom script to format it into TSV files.

Then…

We applied the new TSV files to train a script based on HuggingFace BioBERT model, previously utilized on Barilla ingredient data. The script computed the prediction accuracy by using the provided data.

Finally…

We developed hyperparameter tuning scripts using the HyperBand and Population-based training algorithms from RayTune to predict the accuracy of our entity classifications.

OUR DATA

We used Harvard's n2c2 NLP Research Data Sets which were originally created at a former NIH-funded National Center for Biomedical Computing known as i2b2: 'Informatics for Integrating Biology and the Bedside'. We used the 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records.

HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS

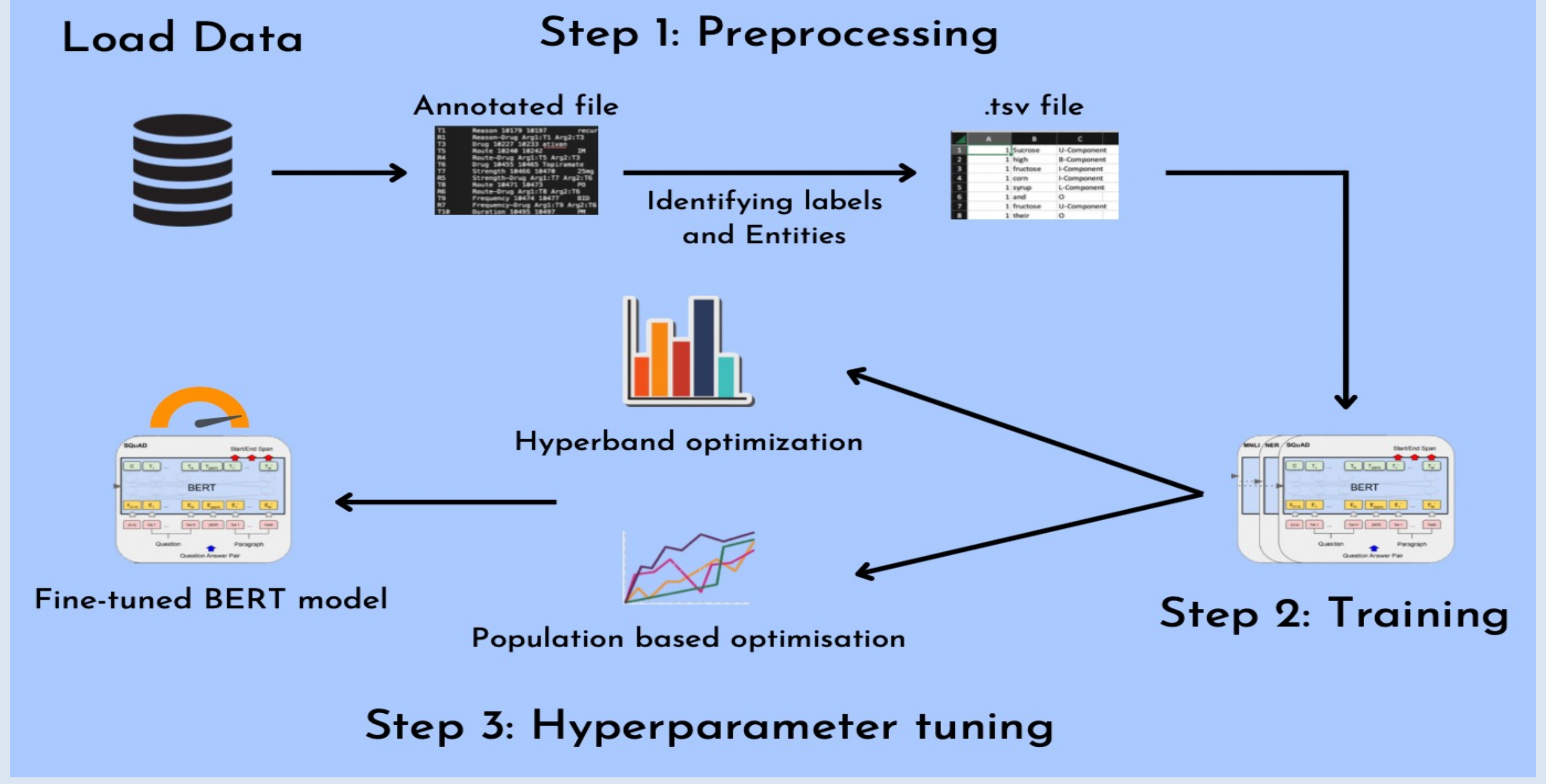
Reference: https://doi.org/10.1093/jamia/ocz166

BERT MODEL

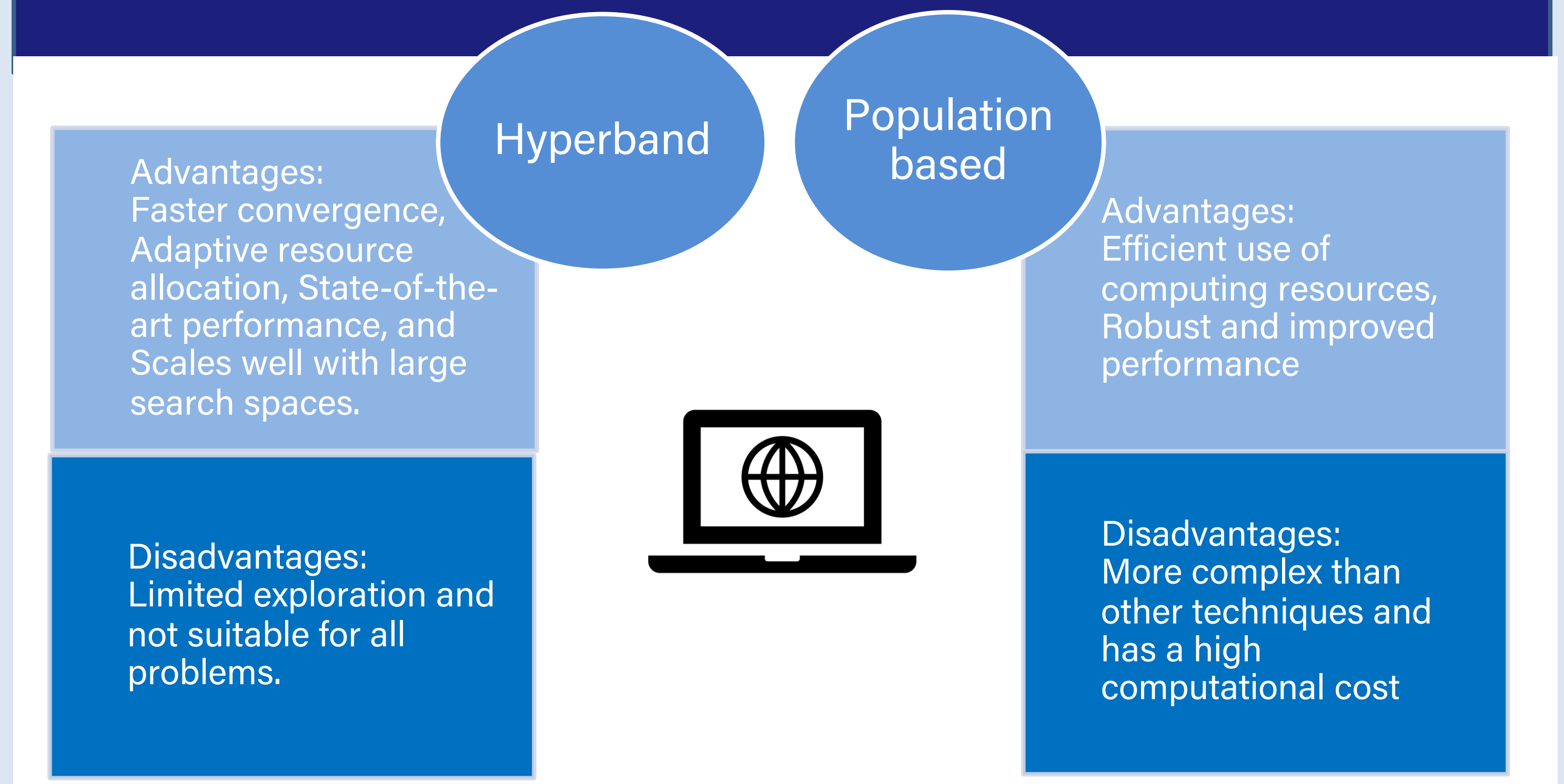
BERT
- Large neural network that can be pre-trained to gain a general understanding of language

Bio BERT
- A variant of BERT specifically designed for biomedical text mining tasks like NER and SR.

TRAINING & FINE TUNING



KEY TAKEAWAYS



Hyperband

Advantages:
Faster convergence, Adaptive resource allocation, State-of-the-art performance, and Scales well with large search spaces.

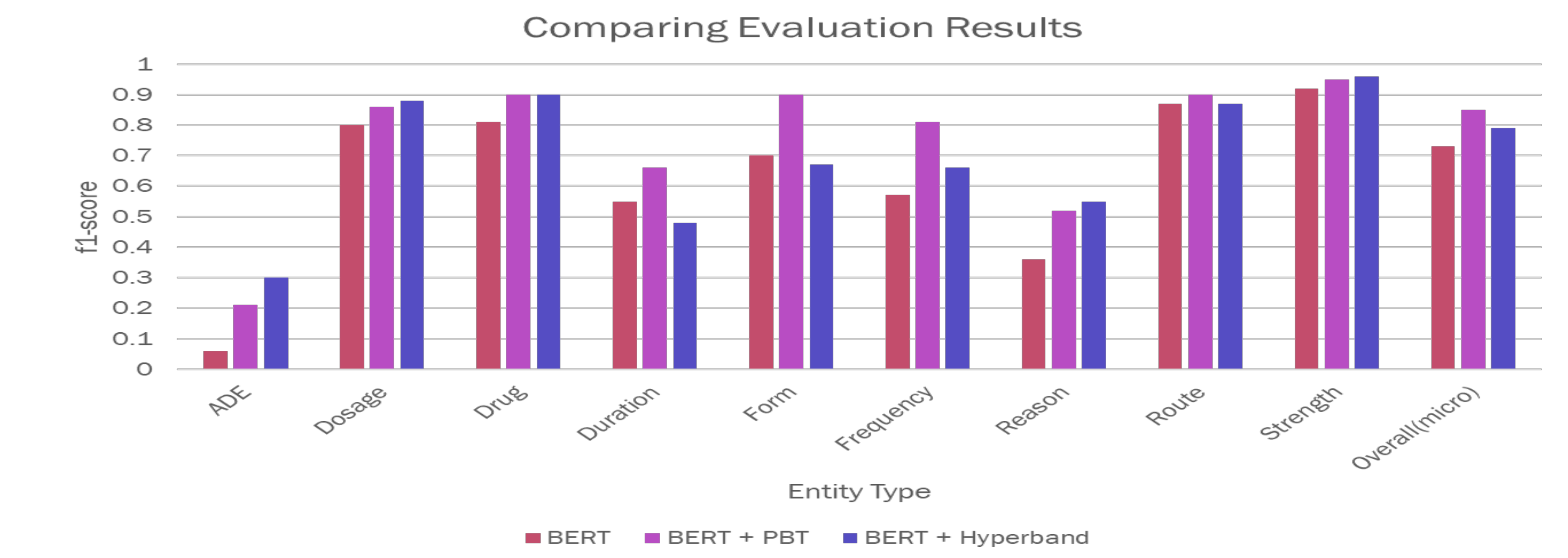
Disadvantages:
Limited exploration and not suitable for all problems.

Population based

Advantages:
Efficient use of computing resources, Robust and improved performance

Disadvantages:
More complex than other techniques and has a high computational cost

RESULTS


Comparing Evaluation Results

FUTURE STEPS & APPLICATIONS

- In the future, we will try to improve the model's performance for every entity category, such as raising the accuracy of the ADE class by labelling more tags in the data.
- We could further expand the scope of the project to include a relation extraction model as well, which would enable us to build a knowledge graph of the entities and their relations.
- We can apply this model to other medical datasets, like those provided by n2c2 and i2b2.

CONCLUSION

- We developed hyperparameter tuning scripts using the HyperBand/Population-Based Training algorithm to tune a BERT model to identify entities from electronic medical records.
- The results after hyper parameter tuning conclude that there has been considerable improvement in the performance of the BERT model.
- Both Hyperband and PBT optimization tunings have similar results on our dataset, but they come with their own pros and cons.
- However, the model struggles in identifying certain classes of entities like ADE and Duration due to their relatively lesser number of labels, but we plan to improve on this area further in our research.

ACKNOWLEDGEMENTS

A very special thank you to our TAs Ujjwal Garg and Yuhang Fang and our Battelle mentors Allen Chen and Mitch Gauthier. We would also like to thank Dr. Ward, Margaret Betz, and the entire Data Mine staff for their continuous support.

Dwijen Chawra - Vamsi Deeduvanu - Aryan Samantaray - Shivli Agrawal - Fengxu (Kenny) Liu - Carissa Lukac - Nitin Murthy - Dave Patel

Google AI | DMIS | HARVARD MEDICAL SCHOOL