

# 1 Natural Language Processing

In this chapter we introduce modern NLP libraries, techniques and their applications. This chapter will focus on deep learning methods and less on computational linguistics that require nuanced knowledge of linguistics. We explore what it means to represent words and sequences of words with rich numeric representations that are better-suited toward modern computational tasks. We aim to capture some of these modern fine-tuned representations that are specially catered toward a semantic lexicon for medical language. We use these representations and aforementioned tools to showcase a modern reference implementation leveraging PyTorch, PyTorch Lightning and the Huggingface Transformers library. To wrap it all together, we walk through a complete example that highlights best practices that encourage reproducibility and allow for systematic iterative improvements.

This includes:

- Bootstrapping techniques to iterate on a dataset in the low-resource setting
- Storing of a reference dataset in a publicly-accessible location
- Downloading, caching, loading, splitting, and preprocessing of the data
- Setting up of a cloud-based GPU workstation (?) (–this might be overkill for now, but keep if we can)
- VSCode (?)
- Monitoring the training run:
  - Logging and experiment tracking
  - Learning curves
  - Metrics
- Hyperparameter tuning, some tricks of the trade
- Offline evaluation and sanity checking

We will keep the discussion focused on SUDEP prediction from electronic medical record (EMR) notes. Many of the concepts introduced here are very general and are straightforward translations to domains outside of SUDEP prediction, epilepsy, and even NLP.