Lecture 5

QM 701: Advanced Data Analytics

Fuqua School of Business

2024

Guest Lecture for Class 6

- First ~35 minutes of the session
  - The talk is pre-recorded and Professor Si will be there live to answer questions
- Lecturer: Professor Shijing Si
- Brief Bio
  - Associated Professor at Shanghai International Studies University
  - Postdoctoral research @ Duke from 2019-2020
  - Research in the intersection of NLP, machine learning, and healthcare.

Administrivia

- The Homework 5 file was updated on Friday
  - If you started HW 5 before Friday, you do not need to redo it. There are just several minor changes in the wordings of some questions.
  - Submissions for HW 5 will be closed after 11 am on Aug 10th.
- Logistics of the Final Exam
  - You can take the exam any time between August 10th 2 pm and August 19th 11:59 pm. I will post HW 5 solution at 11 am on Aug 10th
  - The exam will take 2 hours, it is consisted of multiple choice and short answer questions.
  - If you want to start early, start reviewing all homework solutions and make sure you understood them.
  - More resources for preparing the exam will be posted before August 10th

Language Models

A Brief Timeline for Language Models

1. N-gram (Module 2)
   - The simplest language model, predicts the next word using the previous N-1 words based on counts
2. Feedforward Neural Network (Module 4)
   - Uses a neural network to predict the next word based on the embeddings of the previous N-1 words
3. Recurrent Neural Network (Module 4)
   - Maintains a hidden state in each time step that summarizes the previous inputs, hence allows the model predict the next word based on all of the previous words
4. LSTM Recurrent Neural Network (Module 5)
   - Maintains both hidden and cell states, the states are updated through gating mechanisms to better capture long-range dependencies
5. Transformer (Module 6)
   - Uses multiheaded self-attention mechanisms to capture long-range dependencies
   - The transformer models can be trained in parallel on GPUs, leading to state-of-art large-language models

Causal and Masked Language Models

Causal Language Models

- Training Objective: predict the (likelihood of) next word based on the previous words
- Context: from left-to-right
- Examples: GPT, Gemini, Llama, etc.
- Applications: text generation

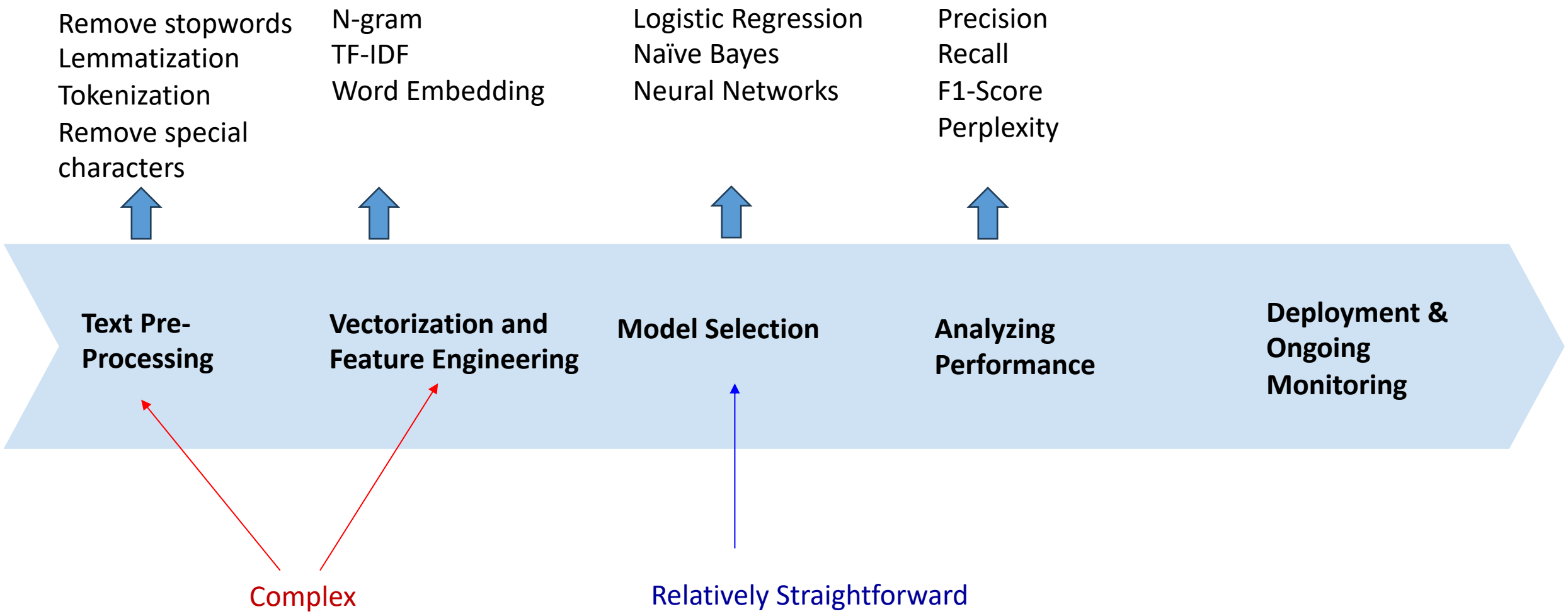
Masked Language Models

- Training Objective: predict the masked words based on the surrounding words
- Context: bidirectional
- Examples: BERT, RoBERTa, ELMo, etc.
- Applications: text comprehension

Both causal and masked language model outputs contextual embedding, i.e., a numerical vector that representing the contexts of the given words. For masked language models, the contextual embedding is often more important than predicting the masked words.

Finetuning

Pipeline from Classes 1 and 2



Pipeline for Finetuning BERT

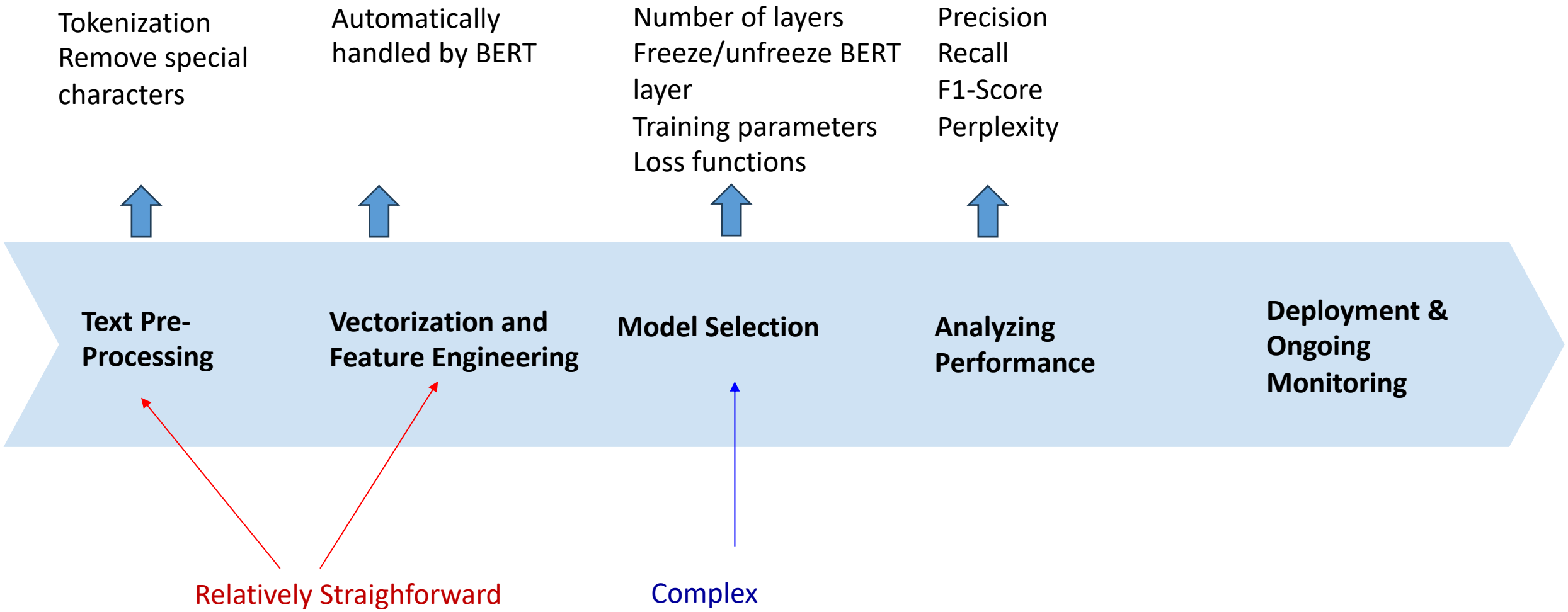
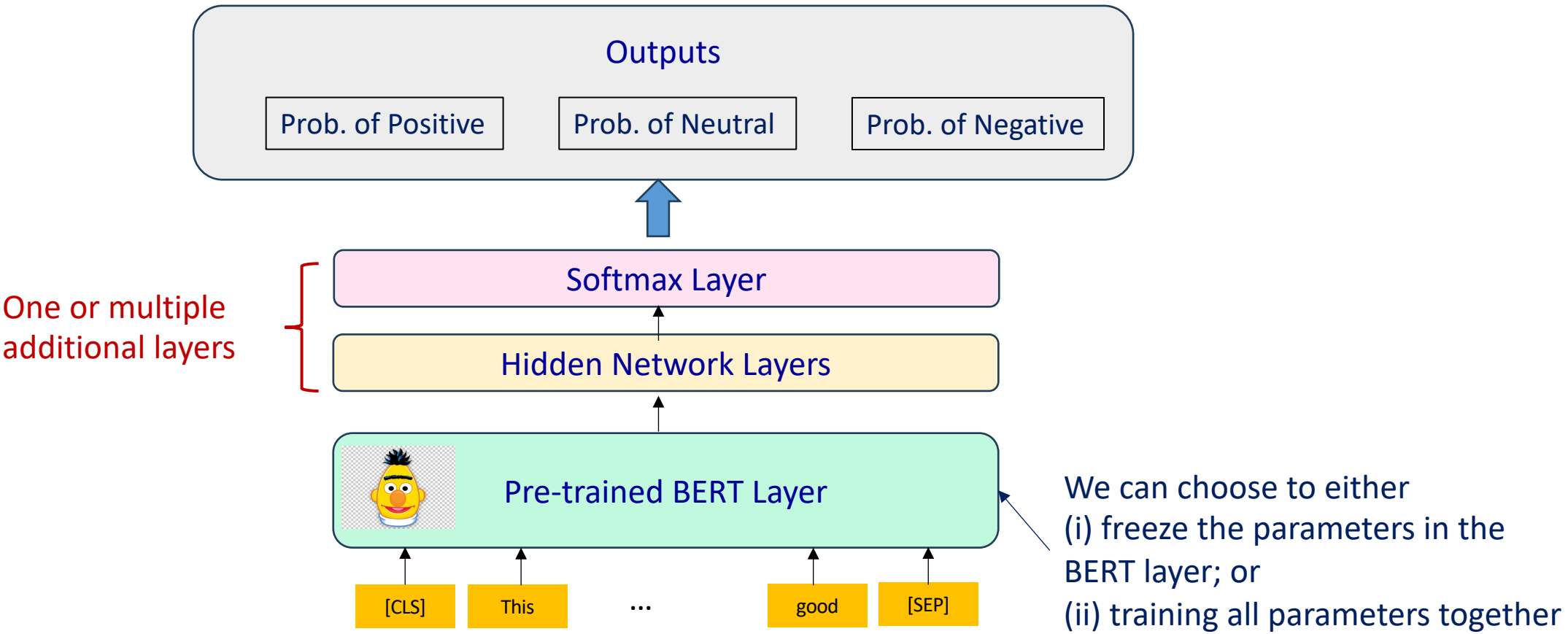


Illustration of a Finetuned BERT Sentiment Analysis Model



Homework 5

Finetuning BERT for Financial Sentiment Analysis

- Q1: Inspecting and Splitting Dataset. (15 points)
- Q2: Establishing Benchmarks (20 points)
- Q3: Classification with the Pre-trained BERT Model (15 points)
- Q4: Finetuning the BERT Model (50 points)
- Bonus: Sentiment Analysis with FinBERT (10 points)