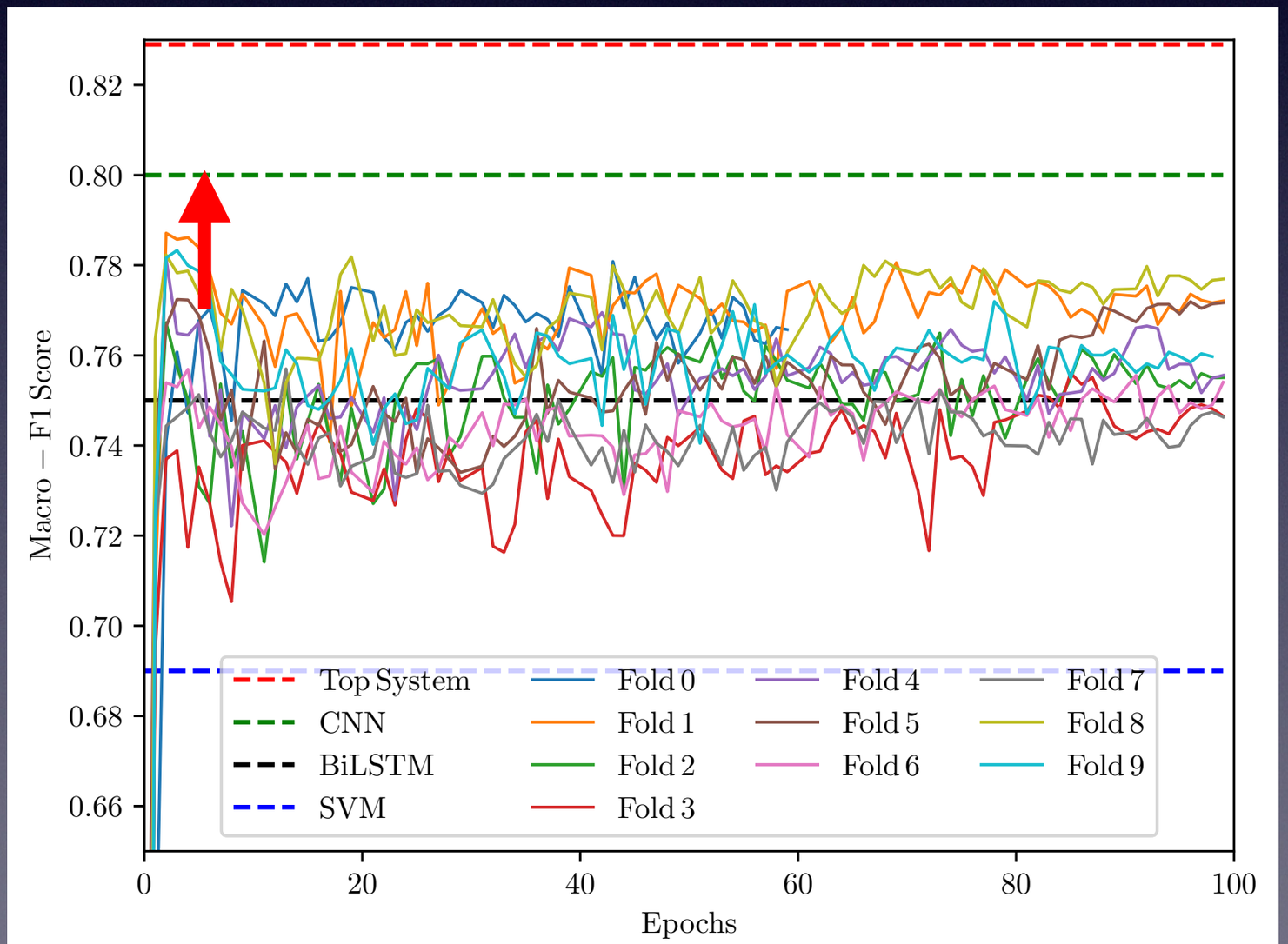
Fine-Tuned BERT Ensembles for Offensive Language Identification

Jennifer Kadowaki (jkadowaki@email.arizona.edu)

HOW DO YOU PICK THE BEST EPOCHS TO ENSEMBLE?

- **Experiment #1**: ***The best individual performers***
- **Experiment #2**: ***Last n epochs with and w/o early stopping***
- **Experiment #3**: ***Least redundancy***



Fine-tuned BERT models for ~60 GPU hours producing 1000 BERT checkpoints totaling ~1.5 TB!

You don't need to go to such extremes to get a good ensemble!

Ensemble	
Epochs	2,3,4
Macro-F1 Score	0.800512
Accuracy	0.843023
Weighted-F1 Score	0.841203

***Table 1*: Final Results of Task 6a.**