Parnal Sinha L06: AWS MLU Module 2 Reflection
ITAI 2376: Data Science in AI
Date: 05/08/2025
Lab Module: Module 2 – Natural Language Processing & Transformers

AWS MLU Module 2: Natural Language Processing

Reflective Journal

Learning Insights

Throughout Labs 1 to 5 of Module 2, I developed a comprehensive understanding of natural language processing (NLP) and deep learning techniques for text data. The labs systematically introduced me to increasingly sophisticated NLP workflows – from basic text cleaning and visualization to deep learning-based language modeling using BERT.

The early labs reinforced fundamental concepts such as tokenization, stemming, lemmatization, and named entity recognition (NER), while also teaching important practical techniques like part-of-speech tagging and vectorization (BoW, TF, TF-IDF). These concepts clarified how raw text data can be transformed into meaningful numerical representations, bridging the gap between human language and machine learning algorithms.

Progressing into Labs 3 and 4, I explored **word embeddings** using GloVe vectors and implemented **recurrent neural networks (RNNs)** for text classification. These experiences deepened my understanding of how models capture semantic relationships and temporal patterns in text.

The final lab was particularly impactful. I fine-tuned a **DistilBERT** model on sentiment classification tasks, gaining hands-on experience with transformers and transfer learning. This not only reinforced my technical understanding of attention mechanisms but also provided exposure to the challenges of working with large pre-trained models, such as managing memory and computational constraints.

One of the most valuable learning moments was recognizing how **incremental complexity** – from basic vectorization to advanced contextual embeddings – reflects real-world NLP pipelines. The labs provided a microcosm of how modern NLP solutions evolve from simple text representations to leveraging state-of-the-art language models.

Challenges and Struggles

Several technical challenges emerged during the labs. Early on, handling text preprocessing nuances (stop words, stemming/lemmatization inconsistencies) required careful experimentation. In Labs 3 and 4, managing word vector dimensions and aligning embedding matrices posed difficulties, especially when integrating GloVe embeddings into RNNs. However, due to working on the midterm and other workflows, the fixes and errors were relatively easy to spot and rectify quickly.

Parnal Sinha L06: AWS MLU Module 2 Reflection
ITAI 2376: Data Science in AI
Date: 05/08/2025
Lab Module: Module 2 – Natural Language Processing & Transformers
The most significant challenge occurred in Lab 5 while fine-tuning DistilBERT. I encountered **library version mismatches** between `transformers` and `torch`, specifically regarding new features like `torch.compiler`. Resolving this required downgrading the `transformers` package to maintain compatibility with my environment. This issue not only tested my debugging skills but also reinforced the importance of environment management when working with cutting-edge NLP tools. Again, a far cry from the errors in the Diffusion model assignment but still a worthwhile problem solving in the labs to ensure I recognize these simplistic runtime errors.

Training large models also stressed hardware limitations, requiring me to adjust batch sizes, monitor GPU memory, and use checkpointing strategies to avoid runtime crashes.

Through these obstacles, I developed stronger problem-solving strategies, including:

1. Proactive version control and environment isolation.
2. Iterative debugging using minimal examples.
3. Strategic use of model freezing and batching to reduce compute overhead.

Personal Growth

Module 2 marked a significant leap in my understanding of **deep learning for text data**. I transitioned from viewing NLP as a preprocessing task to appreciating it as a layered process involving both **linguistic theory and computational modeling**.

One area of growth was mastering **vector space thinking**: understanding how words and sentences can be represented and manipulated mathematically. This has already informed my broader research interests, especially in designing AI agents for text classification, media understanding, and even biomedical NLP applications.

I also gained patience and confidence in navigating advanced ML workflows, particularly when experiments did not yield immediate results. This resilience will be invaluable as I pursue more ambitious projects involving transformers, RAG (retrieval-augmented generation), and multimodal AI systems.

Critical Reflection

If I were to repeat these labs, I would invest more time in:

1. **Experimenting with hyperparameters** for both the RNN and BERT models.
2. Exploring alternative vectorization strategies like Word2Vec and FastText.
3. Integrating additional evaluation metrics beyond accuracy and loss, such as F1 scores and confusion matrices.

Moreover, I would attempt to scale the models to larger datasets and investigate **advanced transfer learning** techniques, including unfreezing encoder layers for fine-tuning in Lab 5.

Parnal Sinha L06: AWS MLU Module 2 Reflection
ITAI 2376: Data Science in AI
Date: 05/08/2025
Lab Module: Module 2 – Natural Language Processing & Transformers
Looking ahead, I am eager to extend these NLP skills into **real-world applications**, such as sentiment analysis, document summarization, and AI-driven customer support systems. The Module 2 labs have provided a solid foundation for both academic research and industry-focused NLP development.

Finally, these labs reinforced the broader significance of **machine learning pipelines** – from data preprocessing to model evaluation – which mirrors the best practices used in professional AI development environments. Regardless, these notebooks will be quintessential in the future if I ever wish to rerun the workflow – the methods can simply be interpolated when necessary for the specific use cases. The lab was crucial and informative for further work in AI and ML analysis.

References

1. AWS Machine Learning University Module 2 Labs and Documentation
2. Hugging Face Transformers Documentation
3. PyTorch Official Documentation
4. Scikit-learn Documentation
5. Course Lecture Materials