WIDS-2025
Project Report week3

Boda Prabanjan Jadav
24B1001
Mentor : Aditya Sanapala

Contents

1 Introduction

In this week, I worked on understanding and applying modern Natural Language Processing techniques based on Transformer architectures. The focus was on using pre-trained transformer models for inference, training, tokenization, and text summarization. Through hands-on coding tasks, I learned how large language models are practically used in real-world NLP applications.

2 Understanding Transformer-Based Inference

I learned how to use pre-trained transformer models directly for inference using high-level pipelines. By applying sentiment analysis on example text inputs, I understood how models can be loaded and used without manually defining network architectures. This helped me appreciate the power of transfer learning and how pre-trained models generalize across tasks.

3 Working with Tokenizers

I explored how tokenization works in transformer-based models. I learned that text is not processed as raw words but is broken into subword units using tokenizers. This helped me understand why tokenization is critical for handling unknown words and maintaining consistency across different languages and domains.

4 SentencePiece and Subword Tokenization

I learned about SentencePiece as a language-independent tokenization method. By loading and using a trained SentencePiece model, I understood how text can be converted into subword tokens without relying on whitespace. This clarified how modern NLP models efficiently handle vocabulary size and rare words.

5 Using Datasets for NLP Tasks

I learned how to load and work with standard NLP datasets using the datasets library. This helped me understand how large datasets are structured and how they can be easily integrated into training and evaluation workflows. I also learned how datasets are split into training and validation sets for model development.

6 Mini-Project: Text Summarization

In the mini-project, I used a transformer-based summarization model to generate concise summaries from long text passages. This helped me understand how encoder-decoder transformer models work for sequence-to-sequence tasks. I observed how the model captures key ideas while reducing text length, which demonstrated the practical usefulness of transformers in real applications.

7 Key Takeaways

From this week, I learned:

- How transformer models can be used directly for NLP inference tasks.

- The importance of tokenization and subword representations.
- How SentencePiece enables language-agnostic tokenization.
- How to use standard NLP datasets efficiently.
- How transformers perform complex tasks like text summarization.

8 Conclusion

Overall, this week strengthened my understanding of modern NLP workflows using transformers. I gained both conceptual clarity and practical experience, which helped me see how state-of-the-art NLP systems are built and deployed. This foundation will be useful for more advanced tasks involving fine-tuning and building custom NLP models.