Research Assessment

Kavan Mehta

Ms. Whitcomb

ISM 2

18 November 2022

Kavan Mehta

Ms. Whitcomb

ISM 2

18 November 2022

Research Assessment #12 (Original Work Related)

Date: 18 November 2022

Subject: Hugging Face Behind the Pipeline Lesson

MLA citation(s):

"Behind the Pipeline - Hugging Face Course." *Hugging Face Course*, Hugging Face, https://huggingface.co/course/chapter2/2?fw=pt.

Assessment:

Currently, I have been exploring the Hugging Face library by viewing tutorials and online documentation to understand the different intricacies that I need to use to develop a project with transformers in text generation and text classification. While I have been studying algorithms to first develop a foundational understanding of natural language processing (NLP), I also understand the importance of learning the practical implementations of transformers by performing NLP projects. To learn more about the practical implementations with the programming of transformers, I found an educational article from the Hugging Face library itself, "Behind the Pipeline - Hugging Face Course," which went over the science behind pipelines including preprocessing, tensors, and post-processing of transformers.

Through the article from the Hugging Face Course, I first learned about how preprocessing with tokenizers works in transformers using the Hugging Face library. I found out about preprocessing and how tokenizers perform the work of "splitting the input into words,

subwords, or symbols…that are called tokens" and "mapping each token to an integer" to serve use for the model ("Behind the Pipeline - Hugging Face Course" 2). Using the theoretical understanding from my previous secondary research, this relates to how I can implement tokenization with the programming snippets in the tutorial in the Hugging Face Course. I still want to discover more about the mathematics behind tokenization that enables the use of a loss function to increase the accuracy of NLP in major applications such as text classification, text generation, and even sentimental analysis. I also learned about specific programming classes and methods in the Hugging Face library such as the "AutoTokenizer class and its from_pretrained() method" which help me understand how I can use the feature of tokenization to preprocess inputs for my transformer models in my Original Work ("Behind the Pipeline - Hugging Face Course" 2). I will use this information in my code to build my project using Hugging Face with effective tools and libraries. I learned that like other algorithms, Transformers "only accept tensors as input" and "can be scalar (0D), a vector (1D), a matrix (2D), or have more dimensions" ("Behind the Pipeline - Hugging Face Course" 3). This connects to my knowledge from last year that tensors are extremely similar to Numpy arrays in the fact that they were used as the only form of input on the TensorFlow platform for all of my ISM 1 projects. The Numpy vectors that I used in my Final Product and some of my Original Work projects also consisted of multidimensional arrays, which helps me understand the concept of tensors comfortably. Additionally, I gained insight into the aspect of the post-processing of transformers in the learning process. I discovered that some outputs "don't necessarily make sense by themselves," instead they need to be converted to probabilities through a "SoftMax layer…with actual loss function, such as cross entropy" ("Behind the Pipeline - Hugging Face Course" 6). As I used SoftMax functions last year in making projects for computer vision, I can utilize the same

knowledge towards obtaining probabilities with the use of post-processing to make sense of the outputs provided by transformers. This was something essential to reflect upon for my projects with transformers as I need to be prepared on understanding my output using code and utilize this information to make a fully functional solution to a real-world problem.

With this initial step into learning about the practical implementations of transformers through code, I hope to continue to learn about NLP technology through these scholarly articles and research interviews with professionals in the field. I will continue to use the tutorials and the Hugging Face course in order to create Original Work projects that will help me gain a fundamental understanding of my Final Product. Moreover, with a strong theoretical foundation from my previous research in Original Work, I will make important connections between multiple NLP algorithms to facilitate the implementation of these algorithms in ISM 2 to create my Original Work that could help the community.