**Project Title**: Investigating dimensional directions controlling for BERT

**Contributing Students**:

* Tamir Kashi
* Yulia Louzon
* Adva Helman

**Description for the Project**:

In the paper we are going to consider the transformer architecture and the method purposed in the paper GANSpace: Discovering Interpretable GAN Controls in order to discover ways to alter the special [CLS] token, as a mean to changing the style of a textual input while preserving it’s original content. For that cause we will define content as the overall semantics of the text, and style as the unique rhyme and form of the text, the elements that provide it a special feeling.  
StyleGAN is a neural architecture presented at A Style-Based Generator Architecture for Generative Adversarial Networks, providing the ability to change style while preserving content for graphical inputs. We would like to apply similar concept for the encoder and the decoder of the transformer architecture, making it possible to change the style of an input text.  
As done in GANSpace, we would like to use PCA in order to alter the vectors of the last layer of the encoder, moving them in different directions in the vector space in order to achieve different styles when inserted as inputs (along with the original word tokens) for the decoder.  
We are interested in applying those concepts of the graphics field in our domain of NLP, and we are hoping to investigate the involved architectures and learn as much as possible while doing so. There are numerous papers regarding text style transfer, and we would like to contribute to that effort by providing a method that is not based on prompts, but on PCA. The method we suggest has its downsides, like not being able to change the input text into any desired style, but we think it might be interesting to explore and investigate, while also offering a unique approach.

**Requirements:**

We will need access to existing models of textual encoders and decoders. We would need to provide the encoder with enough samples in order to capture the dimensional representation of the embeddings of the [CLS] token.