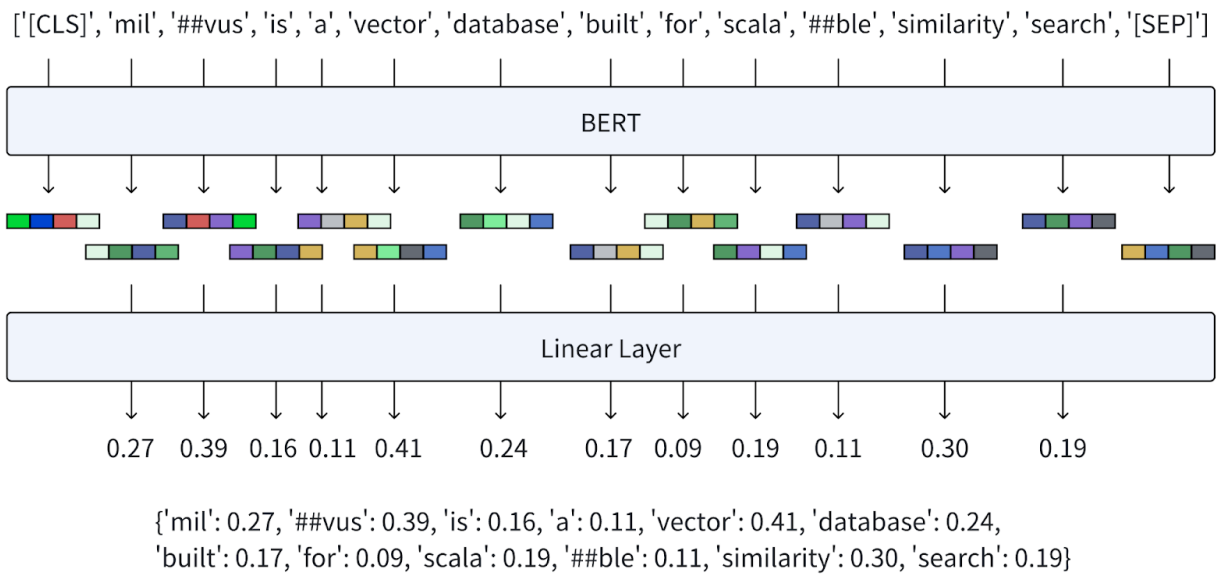
BGE-M3 embedding model

This model is an extension from the capabilities of BERT model, which has produced dense embeddings from the text inputted into the model which has improved quality responses based on the semantic meaning and interrelations within the words/tokens contained in the embeddings, and is responsible in the creation of learned sparse embeddings that balance meanings with good precision which is mostly preferred for nuanced information retrieval.

The contrast difference between bert and bge-m3 model is the more fine approach on capturing the significance of each token:



1. *Unlike relying on [CLS] token representation as in the dense embeddings this model evaluates contextualized embeddings of each token within the sequence where as to signify the importance of token relatively in the sentence.*
2. *An extra layer is added to the stack of the encoder layer from the bert model and now this layer encompasses weight of each token and by passing embeddings through this linear layer bge-m3 obtains a set of weights.*
3. *A rectified linear unit activation function is applied to product of the weights obtained through the linear layer and contextual weight of each token to get*

term weight for each token, this usage of activation function ensure that term weight is non-negative thus enhancing the performance of the sparsity in the embeddings.

4. *The output result is a sparse embedding, where each token is associated with a weight value, indicating its importance in the context of the entire input text.*

This generation of sparse embeddings in this method helps in attaching the importance of each term weight to that of the interrelations where semantic and lexical meanings are considered crucial in search and retrieval in large databases. Thus leading to more precise and efficient mechanisms for searching through and making sense of vast textual data to that of the text used.