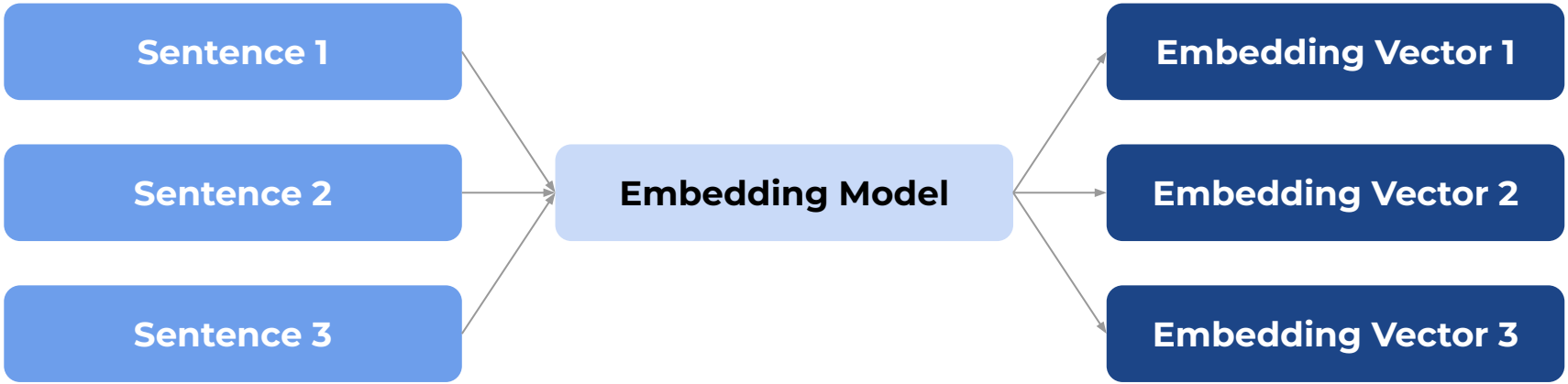
LVC 9: Semantic Search with Transformer Embeddings

Generative AI for Natural Language Processing

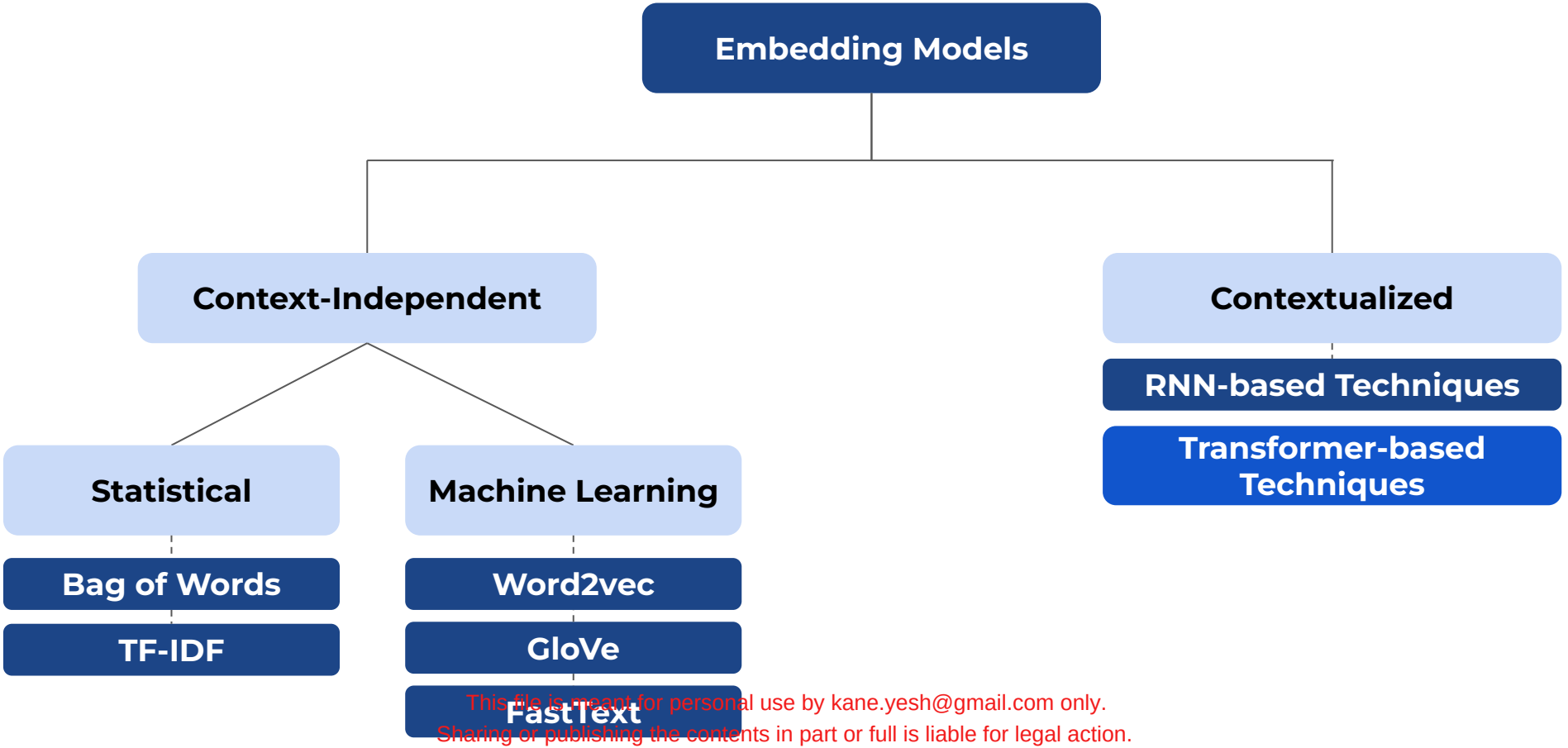
Embeddings as a Vector Representation of Text

Embeddings are lists / vectors of numbers used to represent Natural Language, as they capture the semantic meaning of that text in some multi-dimensional vector space known only to the Neural Network.



Although there have been a variety of methods to generate embeddings for text, researchers have found out that training a Deep Learning model to generate its own embeddings is often the best approach.

Text Embedding Schemes in NLP - A Broad Overview



Transformers - The best way to get Text Embeddings

Further, out of all the Deep Learning architectures researched in NLP, the Transformer model has proven to be the best choice so far to extract high-quality, semantic embeddings for text, for multiple reasons.

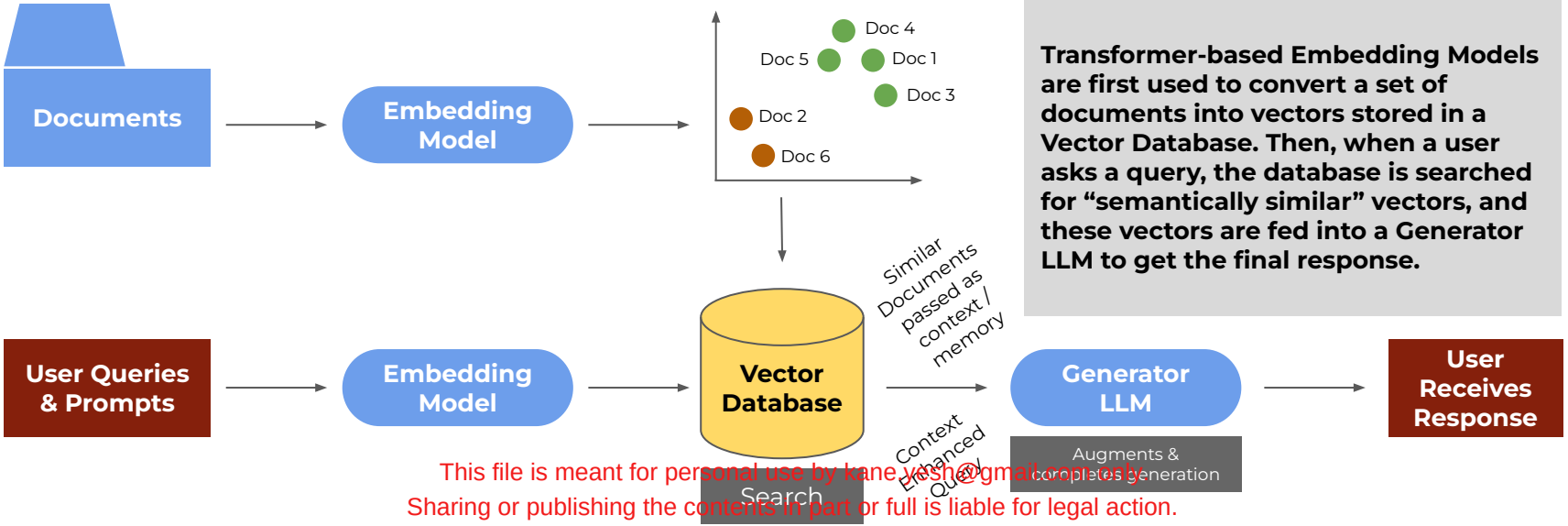
1	**The Self-Attention Mechanism**	Allows the Transformer to simultaneously attend to different parts of a sentence and weigh their relative importance, helping detect contextual relationships among words and long-range dependencies in text.
2	**Positional Encoding**	An additional innovation to add context to the mechanism for computing embeddings, since the position and order of words in a sentence is often critical to its final meaning and understanding inter-word relationship.
3	**Sentence Embeddings**	Because words in language can change their meaning depending on the sentence & other words they're surrounded by, it's critical to get not just word embeddings but sentence embeddings for true semantic understanding.
4	**Parallelized Training Procedure & Huge Text Corpuses**	Due to the parallelized training architecture of Transformers, they can be trained faster by multiple GPUs and their training can scale faster to large text corpuses, allowing them to also arrive faster at higher quality embeddings.

For all these advantages, Transformer-based Embedding models & Large Language Models are now considered the de facto standard in vectorizing text and encapsulating semantic meaning in a useful way.

Embeddings for Semantic Text Retrieval

One of the most useful applications in NLP for these text embeddings, is "Semantic Search" - the ability to understand the semantic meaning of the user's search query, and retrieve / return only those results that match this semantic meaning to some level of similarity. While the idea has been around for long, it has only recently become possible due to the high-quality semantic embeddings created by Transformers.

"Semantic Search" is the main tool behind the idea of RAG - Retrieval-Augmented Generation.



Transformer-based Embedding Models are first used to convert a set of documents into vectors stored in a Vector Database. Then, when a user asks a query, the database is searched for "semantically similar" vectors, and these vectors are fed into a Generator LLM to get the final response.

Similarity Search Demonstration