**RAG-assisted explainable hate speech detection model from GenerativeAI**

In our proposed work, we have compiled a corpus sourced from Twitter, encompassing five distinct topics: sports, entertainment, politics, religion, and the economy. The project's primary objective is to construct a Retrieval Augmented Generation (RAG) model for fine-tuning a Language Model (LLM) on our designated corpus to predict hateful comments with coherent reasoning.

The functionality of the proposed model unfolds in two main phases:

**Retriever Part:** the first part is to build the retriever component by extracting relevant blogs and articles associated with the collected tweets. These extracted articles undergo a rigorous cleaning and decomposition process before being fed into encoder models to generate embeddings. These embeddings are then loaded in a vector database for subsequent semantic searching.

**Generator Part:** The second part of our model involves accepting user queries, where the queries are tweets collected in Hindi-English code-mixed text. Initially, this code-mixed text is translated into monolingual English and subjected to further cleaning and encoding processes. The user query's resulting encoded value serves as a semantic search vector. This search vector and an optimal 'K' value are employed to retrieve K semantically similar context information from the vector database.

Subsequently, these K contextual chunks undergo processing through a text summarizer model. The output from the summarizer is then bifurcated into two channels: the first channel acts as an input to the augmentation model. This channel combines the user query with contextual details to prompt the Language Model (LLM). The second channel directly provides contextual details for the predicted label class.

**Technical Exposer: Retrieval Augmented Generator(RAG), Prompt Engineering, Vector Database, LLM, Text summarization, and Text classification.**

