**BERTopic: Neural topic modeling with a class-based TF-IDF procedure**

BERTopic is a neural topic modeling technique with the goal of uncovering common themes and the underlying narrative in text data.

There are three steps in BERTopic.

In the first step, it creates document embeddings using pre-trained transformer-based models. It uses Sentence-BERT (SBERT) to embed documents into a semantic vector space, and these embeddings can be created with any model fined tuned for semantic similarity.

In the second one, it creates clusters of semantically similar documents. The process to do this is to use UMAP to reduce dimensionality of embeddings and then to apply HDBSCAN to cluster.

Finally, in the last step, it generates topic representations. This is done by treating each cluster as a single document and then, using class-based TF-IDF approach (c-TF-IDF), highlighting important words in each cluster.

**A close up of a word

AI-generated content may be incorrect.**

(Formula of c-TF-IDF)

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

AI-generated content may be incorrect.A screenshot of a report

AI-generated content may be incorrect.

The MiniLM model offers a strong balance between speed and performance, making it particularly suitable when computational resources are limited.

The main strengths of this approach are its flexibility, scalability and its dynamic. The flexibility is due to the separation of document embedding and topic representation and the fact that there is no need to re-cluster when fine tuning topic generation. C-TF-IDF allows dynamic modeling of topics and cross-class comparisons.

Although it has strong performance, it also has some limitations when dealing this multi-topic documents and contextual topic representations. The first can be mitigated by the use of HDBSCAN’s probability matrix.