**Create topic model that will extract important topics using BERT**

**Background:**

With the growing amount of data in recent years, that too mostly unstructured, it’s difficult to obtain the relevant and desired information. Topic modeling is a great way of dealing with huge corpus with unlabeled data.It is**a process to automatically identify topics present in a text object and to derive hidden patterns from a corpus**

**Flow:**

Create Embeddings{BERT] ---> Dimensionality reduction[UMAP] --> Clustering[HDBSCAN] --> Topic Creation[Class tf-idf-] --> Extract top words

**Result:**

- Easily interpreted topics using top 20 words

-Sports, computers, space, and religion are few topics that were extracted from the data.

-Reduced 56 topics to 36 using topic reduction

------------------------------------------------------------------------------------------------------------------------------------------

A. Data

20 Newsgroups dataset which contains roughly 18000 newsgroups posts on 20 topics

B. Model:

Topic models such as LDA and NMF are good starting points but require hyperparameter tuning to create meaningful topics. Hence BERT is used

Major advantages of using BERT:

1. Pretrained model available--reduces training time and complexity

2. Embeddings can be extracted based on word context---Gives better word representation

***I Embeddings:***

-BERT is used to get embeddings based on the context of the word

-Sentence-transformers package is used as it works well on doc level

-Used Distilbert as it gives nice balance between speed and performance

***II Dimensionality reduction***

- Vector size is 512 and has to reduced before clustering to get better clustering performance

-We use UMAP to reduce dimensionality as it is proven to contain maximum info with reduced vector size

***III. Clustering using HDBSCAN****(Image below)*

-It is preferred as not all documents are forced towards a certain cluster. If no cluster could be found, then it is simply an outlier.

A picture containing chart, scatter chart

Description automatically generated

***IV Topic Creation***

-We are interested in understanding what makes a cluster unique from others

-Basically derive topics from clusters

-We use class based variant of tf-idf--(c-tf-idf)

-Basically treating all docs in a cluster as one doc and understanding word importance

-Select top 20 imp words based on tf-idf scores . These are most representative of the topic!

***V. Topic reduction***

-Based on data we can get a lot of topics which might not be very useful

- HDBSCAN will get fewer topics through its min\_cluster\_size parameter but it does not allow to specify the exact number of clusters

- So we merge the topic vectors that are most similar to each other.