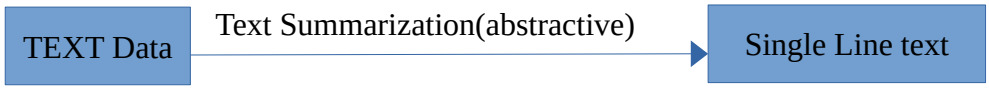
In Social media, all data includes these three types: Text, Video, Image.
We need an agent that will take any of these data types and output an encoded vector in a common feature space so that we can compare encoded vectors of text with encoded vectors of video to perform clustering.
But first we need to make the dataset.

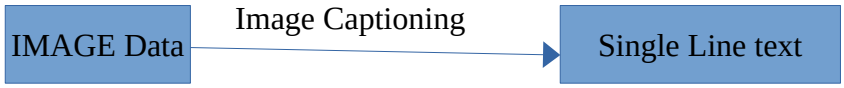
1. Dataset making:
    1. Scraping over the social media, and grab all data of these three types.
    2. Separate them according to these types.
2. Extracting encoded vectors from each type of data to a common feature space.

    a. Text data processing:

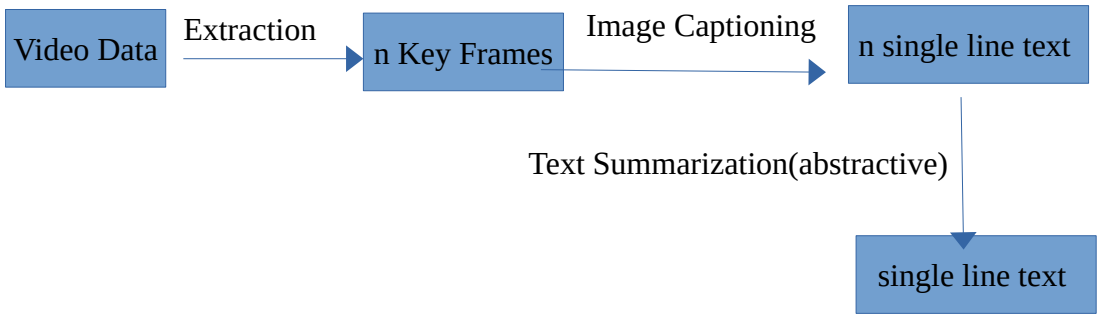


Here, a paragraph of text which is a social media post, will be transformed into a single line.

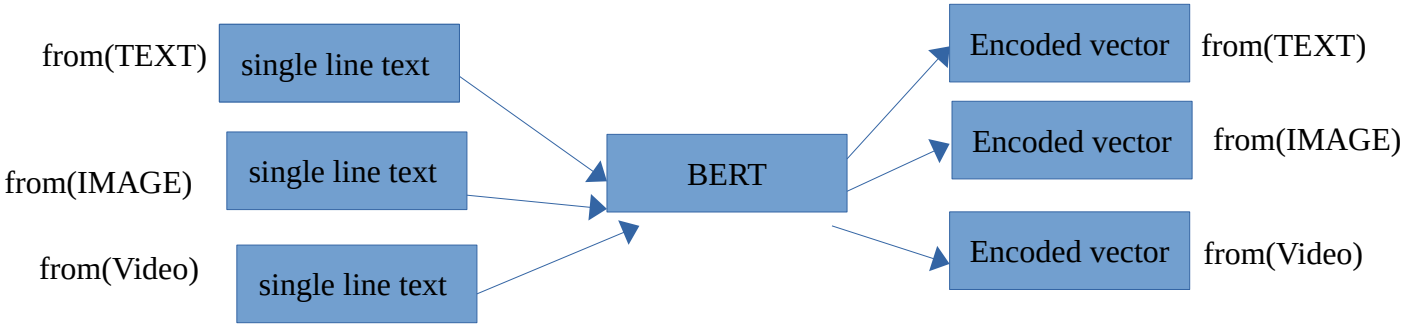
b. Image data processing:



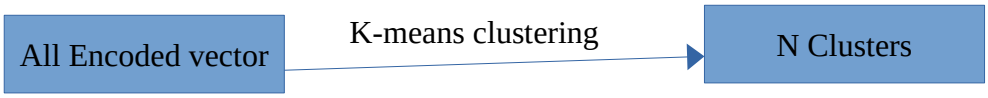
c. Video data processing:



3. All these “single line text” from these three data type will be feed into a model named “BERT” to get encoded vector.



4. These Vector will be feed into K-means model to find out cluster and make assumptions.



Now combining all these sub processes into one single agent, here is a abstract view: