Day 1:

I created a first model for filtering E, S or G using sentence transformers library in python

It’s a multi-ling model “paraphrase-multilingual-mpnet-base-v2”

The idea is to take the average embedding of the dictionaries associated with the E, S and G. Then we compare this “average embedding” with each sentence in our data frame using two methods:

* Euclidian Distance
* Cosine Similarity

Day 2 & 3:

Idea: Take a tweet -> extract keywords -> compare those keywords to the ESG dictionary at hand

We eliminated Euclidian Similarity since our model was pretrained using cosine similarity.

We worked on Preprocessing the tweets in order to extract useful information from the tweet and feed it to our model.

The main work was on taking the keywords from the tweet text and measure its cosine similarity with the ESG dictionary we have at hand.

The results till now were not great.

Diving further in the preprocessing procedure, I extracted other relevant info from the tweet (mentiosn, urls, hashtags, stocks) and I tested the model.

It’s taking way too long because of the long steps we adopted to it.

Day 4: