WEEK 2 SUMMARY

SurfMetrics

FIRST SIMPLE MODEL IMPLEMENTED

On average it gave some relevant results.

Find better preprocessing tools

The process of filtration using FinBERT could be optimized in Topic Modeling

The sentiment analysis could also be optimized by finetuning existing models

TOPIC MODELING

We end up with a list of topics, each containing a set of associated keywords

The extracted topics are much less than the document collection and its vocabulary

It helps understand whether or not the scrapped data is relevant to ESG

BERTOPIC & TOP2VEC

Both are used each in specific cases.
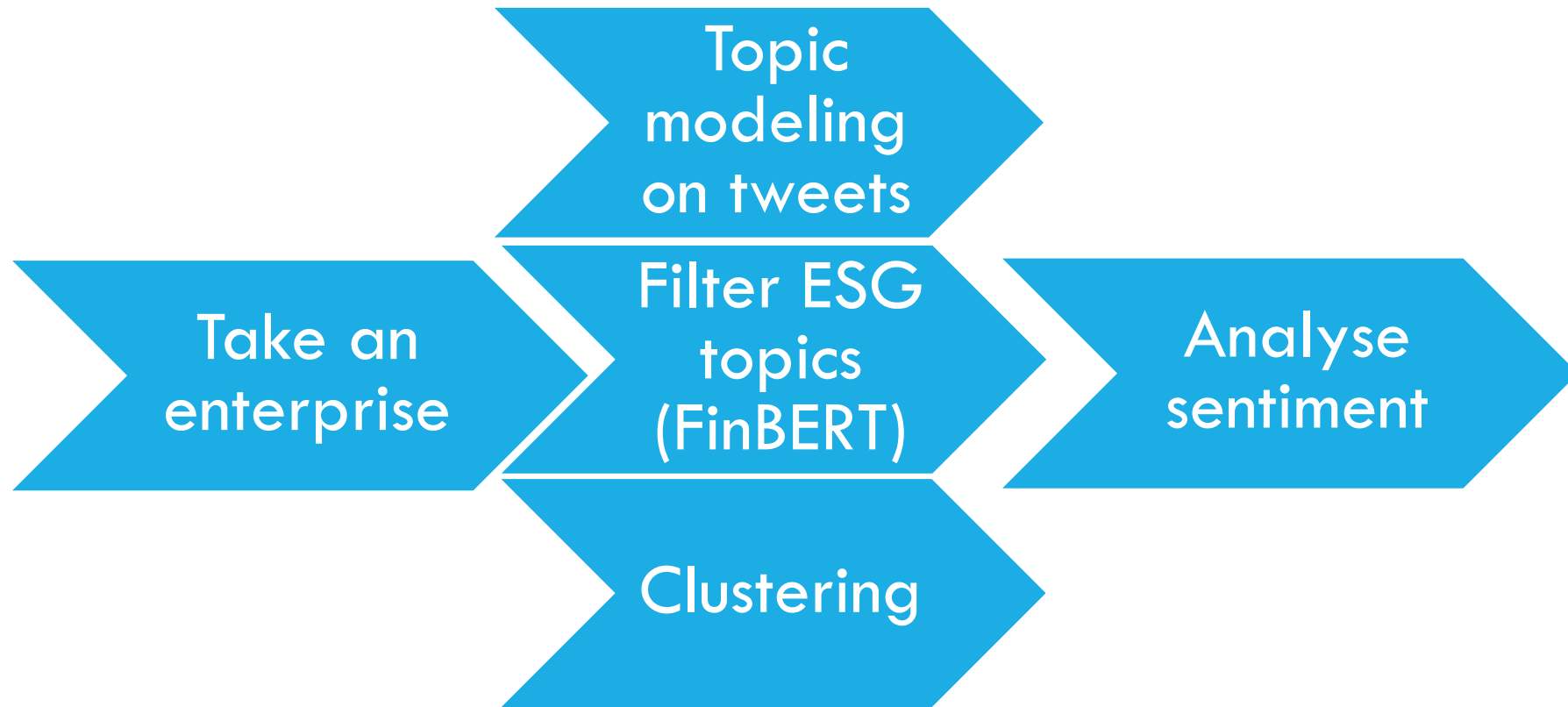
BERTopic was tested for now.

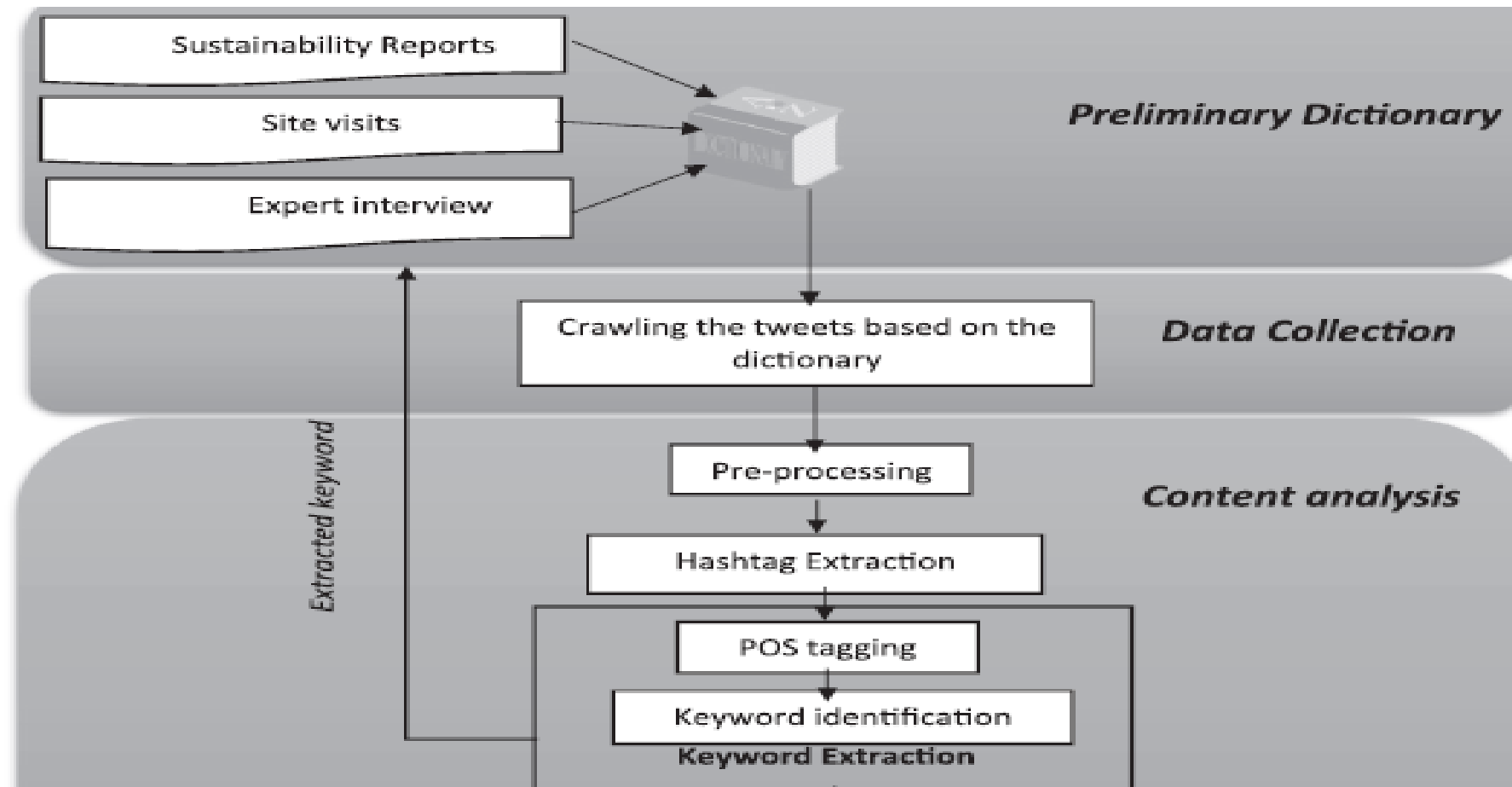BERTopic is more flexible on embeddings nad gave interesting results

CLUSTERING

Tried Kmeans with the Word2Vec Trained embeddings

The results were poor maybe because the scrapped data is not big enough

POSSIBLE VALUE CHAIN

TAKE AN ENTERPRISE

IDEA FOR DISCUSSION

Using the unsupervised finetuning approach to create an embedding capable of understanding ESG

This embedding will serve us for TopicModeling & sentiment analysis

Find ESG texts online → Finetune FineBERT with the same approach as IR → Use the embedding for TopicModeling / filtering