NLP

Problem: Can NLP measure global trade sensitivity at the firm-level from earning calls transcripts?

Idea:

The main idea of this project is to use NLP on firm-level earning calls transcripts to measure the sensitivity of firms to global trade. To "assess" our measure, an event-study can be carried: we predefined periods when events related to global trade happens. Then, we build long-short (long low sensitive – short highly sensitive firms) portfolios and study the abnormal returns around the event. If our measure is a good assessor of firm-level global trade sensitivity, we expect to earn statistically significant abnormal returns during the events.

Step 1: Retrieving data

Earning calls transcripts for US companies are available on Edgar website which is a website on which companies registered at the SEC must comply with some rules namely making financial information (10-K, 10-Q, earning calls transcripts etc.) available in a text or HTML format. The advantages of Edgar are twofold: first, the reporting format is, normally, standardized across companies easing large-scale analysis. Second, these large-scale analyses are made possible thanks to the requirement of HTLM format which allows to web-scrap data. However, the earning calls transcripts can be localized differently on the website across companies and this is a huge problem. CapEdge aims to tackle this issue by providing the information available on the Edgar website in a clearer and better organized manner on their own website. To even further facilitate the data retrieval process, some APIs can help us by providing this information in a few lines of codes (API Ninjas, EarningsCall.biz, ...).

Step 2: Training and assessing the NLP model(s)

The baseline model that can be used is a Bag-of-Word model consisting simply in counting the occurrence of words related to global trade among all the words of the transcripts. This measure is very simple but do not take into account the context in which these words appear, which can be very important in a financial context. New models were developed building on this idea, such as finBERT for example, which is specialized for finance. To assess our model, we will carry an event-study related to global trade and assess the abnormal returns. The best model will be the one that allows an investor to earn the highest abnormal returns related to global trade.

Anticipated challenges:

- Data retrieval and preprocessing, to the extent depending on our choice on how to retrieve the data (web-scraping, APIs)

- Computational time: depending on the coverage (length of the study and number of companies) the training of the model can be long
- Measuring sensitivity to global trade: the main challenge is to take into account the context and if a firm is low or highly sensitive.
- Converting the sensitivity measure into a predictive signal and getting good backtesting results: even if our sensitivity measure is “good” the challenge will be creating predictive signals: when to entry, when to exit, what is the “best” period length for the signals… The portfolio construction step could also be a challenge in the sense that the aggregation of the signals must not dilute its predictive power, if any.

Role of each group member:

- Enzo Montariol: model and event-study
- Matéo Molinaro: model and event-study