

## About the data:

- ❖ Data Source:
  - News data - <https://www.hydrocarbonprocessing.com/news>
  - Price data - <https://finance.yahoo.com/>
- ❖ Data Collection:
  - News data - We have used Beautiful Soup Library to scrape the news headlines from the website with the date information. We have used the news headlines instead of the original news. Since headlines mostly contain the sentiments of the news.
  - Price data - We have collected this data directly from the website.
- ❖ Merged data: We have merged both the news and price data based on the date column.
- ❖ Tenure: Data is collected between Oct 11, 2010 to Feb 02, 2022.
- ❖ Train-Test Split: Data is splitted into 80% of training data and 20% of the testing data.

**Calculating the sentiment scores:** We have calculated the sentiment score using the FinBERT<sup>[1]</sup>. We manually checked that there was a price drop during the first covid wave and the sentiment scores were also negative.

## Reference:

1. Araci, Dogu. "[1908.10063] FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." *arXiv*, 27 August 2019, <https://arxiv.org/abs/1908.10063>.