**BT4222 Time Series Group 5**

| **Item** | **Link** |
| --- | --- |
| Main folder | [Link](https://drive.google.com/drive/folders/1NrTqgTobDwVOvqBAW5LtVapQHqjJlyl2?usp=sharing) |
| Source Code | [Link](https://colab.research.google.com/drive/1gNf84r0hs5GmUd2WCnFmOi5UgX8EZgCj) |
| Data files | [Link](https://drive.google.com/drive/folders/1y3zRkOdx9w0enXPTQWLe7ua1Ubjuyma-?usp=drive_link) |

**Colab files**

1. TS Group 5 source code.ipynb – contains our main source code including:

* Data collection from multiple sources
* Data preprocessing (cleaning + feature engineering + feature selection)
* Exploratory Data Analysis
* ML preprocessing (feature engineering for time series-related variables, train-val-split)
* Traditional machine learning models (XGBoost, Random Forest Regressor)
* Deep learning models (LSTM, GRU, Google Temporal Fusion Transformer)
  + Tested for different train-validation-test split ratio
  + Tested against COVID-19 pandemic from 01-2020 to 06-2020

**Data files**

1. AAPL\_AllNews.csv , AMZN\_AllNews.csv , MSFT\_AllNews.csv, NFLX\_AllNews.csv

* Data collection of news from Polygon.io API, consisting of important information such as news title, date of article published, relevant tickers.
* Sentiment scores and labels were retrieved for each news title from FinBERT.

1. aapl\_processed\_scores\_final\_switched.csv, amzn\_processed\_scores\_final\_switched.csv, msft\_processed\_scores\_final\_switched.csv, nflx\_processed\_scores\_final\_switched.csv

* Data collected from SEC API, consist of quaterly financial reports filling information, and extracted Risk Factor and MD&A sections (2003-2024).
* Data preprocessing (remove section headers, html tags, split into sentences)
* Sentiment scores and labels were retrieved for each sentence from finBERT. This includes sentence counts and sentiment scores for each label in both the Risk Factor and MD&A sections.