**Market Sentiment analysis (News & Stock)**

Target:

* To predict market fluctuation by analyzing the news.
* To quantitative news data. (Based on the Finbert result?)
* Discover the correlation between news and stock.
* HSBC and Bank of East Asia will be tested .
* HSI, HSBC, and BEA stocks price will be downloaded from open-source database.

Assumption:

* News is not the only feature in sentiment analysis.
* Assume each bank should be customized the parameter due to various factors that can affect the accuracy E.g. mother company and subsidiary that can affect their stocks.
* Assume that each news will have multiple reporters discussing the same event, the news should be grouped together and the score of its group should be calculated by standard deviation.
* Assume the model should not include long-term prediction (E.g. More than 7 days), since there will be too many factors affecting the result.
* Assume that news might be leaked out before the media, and the market will fluctuate before the announcement.

Functional Requirement:

* The program should be automated, systematically scrape the latest news, and output the result.
* Multi-Class classification will be used. (Up, Down, Neutral)
* The model should predict the changes in the market within the time range of 3 to 7 days, to provide useful insight for regulators.

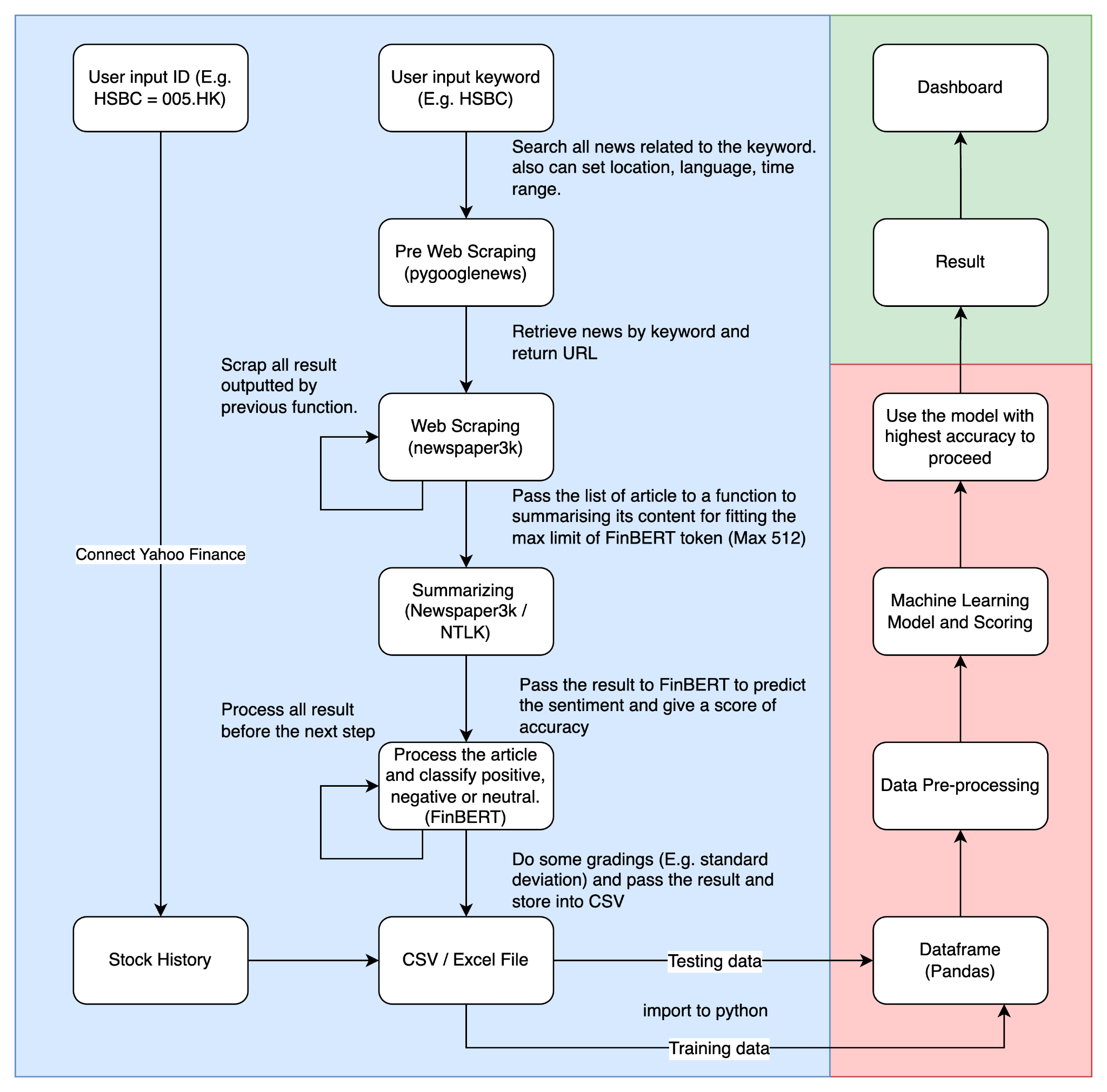
Environment:

* Colab
* Link: <https://github.com/wicksonkold/Market-Sentiment-Analysis>

Package:

| Name | Details | Other Options | Link |
| --- | --- | --- | --- |
| Pygooglenews | Function:  Scrape through Google News by presetting the keyword search.  Pro:  The user can set the time, keyword, and area of the topic (E.g. Business).  Include all news articles online.  Con:  Cannot | Other news API | <https://pypi.org/project/pygooglenews/> |
| Request\_html | Function:  Get all the HTML data by url.  Pro:  Support Javascript that can access platforms like Bloomberg, which Newspaper3k couldn't.  Con:  When requesting the HTML, it requires the user to set a path to extract information from the website. however, different platforms have different architectures, so it would be difficult to scrape all news platforms by using this package. | Newspaper3k | <https://requests.readthedocs.io/projects/requests-html/en/latest/> |
| Newspaper3k | Function:  To read the article by URL and retrieve the main body.  Pro:  Can auto detect the article main body and download it back to python.  Con:  Cannot access the webpage request javascript.  Theres a chance this tools will detect the wrong part of the article and take it as the main body. | Request\_html | <https://newspaper.readthedocs.io/en/latest/> |
| NLTK | Function:  Text Summarization | Pyrrot | <https://www.nltk.org/> |
| Transformer (FinBERT) | Function:  Sentiment analysis | N/A | <https://huggingface.co/yiyanghkust> |
| YFinance | Function:  Access yahoo database and download the data. | googlefinance | <https://pypi.org/project/yfinance/> |

System Architecture



Current Progress:

* Near complete phrase 1 (Area in blue)
* Testing which summeriazer is better compare to others.
* Identifying features.