Current Progress:

* Sentiment (Closed)
  + Completed the automation of fetching data from google news.
  + Finding a way to retrieve news in the past, since google restricted and limited to 100 news per fetch.
  + data will be stored in an excel.
* Summarizer (Closed)
  + BERT (FinBert) has the highest accuracy rate compared to other summarizers (E.g., Spacy, NTLK, and Newspaper3K).
* Identifying features (Hold)
  + News tone score
  + News FLSK score
  + RSI
  + MFI
  + MACD
* Working on TA library and embedding its function to current code.

========================================================================

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

Target:

* To predict market fluctuation by analyzing the news.
* To quantify news data. (Based on the Finbert result?)
* To discover the correlation between news and stock. Cross-correlation should be used to identify the relation between.

Data scope:

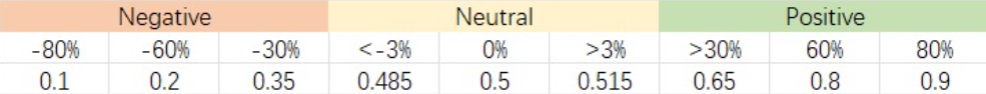
* News: HSBC and Bank of East Asia will be tested
* Stock price: HSI, HSBC, and BEA stocks price will be downloaded from an open-source database.

Assumption:

* News sentiment is not the only feature in predicting stock momentum.
* Including technical analysis in the model will be able to increase the accuracy.
  + Three elements will be included:
    1. Relative Strength Index
    2. Money Flow Index
    3. Moving Average Convergence/Divergence
* Each bank should have its own model with customized parameters due to various factors that can affect the accuracy, E.g., parent company and subsidiary can affect their stocks mutually.
* Each news will have multiple reporters discussing the same event, the news will not be grouped together, because different authors will have different perspectives and the wording will affect the sentiment score. Furthermore, a high amount of negative news discussing the same event can increase sentiment weighting (positive or negative).
* The model should be focused on technical analysis, i.e. fundamental analysis, not including long-term prediction (E.g., More than 7 days) since there will be too many factors affecting the result.
* News might be leaked out before the media, and the market will fluctuate before the announcement. Thus, this might affect the result of cross-correlation.

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 calculated weighting should be classified by the graph below. (Metrics below are just presented as an example, the percentage will be adjusted during further development.)



* The model should predict the changes in the market within a time range, to provide useful insight for regulators. The optimal number of days will be tested with the cross-correlation test.
* the target value of the model should be predicted the next day.

Environment:

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

Package:

| Name | Details | Other Options | Link |
| --- | --- | --- | --- |
| Pygooglenews | Function:   * Scrape through Google News by pre-setting 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 fetch more than 100 news per query due to google constrain. * Originally this library has a date function that allowed the user to fetch data by “From date” & “To date”. But now is not working. | 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's main body and download it back to python.   Con:   * Cannot access the webpage request that supports JavaScript. * There is a chance that this tool 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   Con:   * Some financial terms will be modified. | Newspaper3K  BERT  Spacy | <https://www.nltk.org/> |
| BERT | Function:   * Article Summarization.   Pro:   * Compared to NLTK, Spacy, and Newspaper3k, BERT provides the most accurate result. * Can use a trained model like FinBert to increase the accuracy of the summarization. | Newspaper3K  NLTK  Spacy |  |
| Transformer (FinBERT) | Function:   * Sentiment analysis   Notes:   * “Tone” will be used to process the news sentiment. * “FLSK” should be considered be one of the features to calculate the weighting. | N/A | <https://huggingface.co/yiyanghkust> |
| YFinance | Function:   * Access the yahoo database and download the data. | googlefinance | <https://pypi.org/project/yfinance/> |
| TA (Technical Analysis) | Function:   * To calculate some element of technical analysis. * RSI (Relative Strength Index) * MACD (Moving average convergence/divergence) * MFI (Money Flow Index) |  | <https://pypi.org/project/ta/> |

System Architecture

