

National University of Singapore

School of Computing

Semester 1, AY 2022/2023

BT4222 - Mining Web Data for Business Insights

**Project Proposal**

**Using Machine Learning and Textual Analysis on Macro and Micro Determinants to Predict Stock Returns**

Group: 30

Group Members: A0203270X ASHLEY HO MEI YEE

A0202084U DARRELL LAI WEI AN

A0201534W ELTON TAY JING SHENG

A0199689U WONG YONG DE VICTOR

A0203707M YAP HUI YI

**Table of Contents**

[**Background**](#_yqthfj8kwxfz) **3**

[**Data Overview**](#_mu26ijq76zy) **3**

[Data Sources Description](#_2ksuotcot41m) 4

[**Potential Feature Engineering & Machine Learning Pipeline**](#_rsbvfhxti9dq) **5**

[**References**](#_xp6vauakc0ki) **6**

# 

# 

# Background

Changes in stock prices are difficult to model due to the various macro and micro determinants that affect it.1 Macro determinants can include global and industry level news such as geopolitical tensions, natural disasters, or even speeches by Jerome Powell (Chair of the Federal Reserve of the United States), while micro determinants consist of financial indicators like historical price, profitability, and news on a company level. Most of these determinants are known to influence stock prices,2 and have been used as features in prediction models to predict the changes in stock prices.2 Although, most of these models achieve good results, it is almost impossible to accurately predict the change in stock prices with just historical pricing as global events can heavily affect the prediction model. With the recent market crash, it is even more applicable to create a more resilient prediction model.

Therefore, the problem that our team wants to address is how to create a more resilient prediction model that predicts the daily percentage change in stock prices (henceforth known as stock returns) while maintaining a high performance through unexpected events. We plan to mine insights from news on a global, industry, and company level to give us insights and combine it with advanced machine learning algorithms to develop a more resilient stock returns prediction model.

# Data Overview

This project combines both structured and unstructured data from a variety of sources in order to train a more robust stock returns prediction model. We have chosen the following 4 data sources as they are credible and provide sufficient data points for our analysis.

| **Data Sources** | **Data Types** | **Data Information** | **Collection** | **Shape of Data** |
| --- | --- | --- | --- | --- |
| Google News | Unstructured | Macro Determinants  (News related to War, Jerome Powell, etc.) | Web scraping (BeautifulSoup,  Newspaper3k) | 3 topics of (~1000, 4) |
| Investing.com | Unstructured | Micro Determinants  (Stock Specific News) | Web scraping (Beautiful Soup) | 3 stocks of (~10000, 4) |
| Yahoo Finance | Structured | Micro Determinants  (Historical Stock) | Download (yfinance) | 3 stocks of (1702, 7) |
| US Bureau of Labour Statistics | Structured | Macro Determinants  (Consumer Price Index) | API  (Public Data) | (56,3) |

To ensure that our data has been collected responsibly and ethically, we used public API whenever available. On occasions where we had to scrape data using BeautifulSoup, we made sure to scrape data at a reasonable rate by controlling the number of requests per second with a time.sleep() function in our code. This prevented the scraper from causing the server to crash.

## 

## **Data Sources Description**

To keep the data consistent across all data sources, data are collected from the same timeframe starting from January 2018 to August 2022.

| **No.** | **Unit** | **Type** | **Description** | **Units** |
| --- | --- | --- | --- | --- |
| **News (War, Jerome Powell, Natural Disasters, Stock Specific)** | | | | |
| 1. | news\_title | String | Title of the news article | Title |
| 2. | link | String | The URL to the news article | URL |
| 3. | published | Date | The published date of the news article | Year/Month/Day |
| 4. | article\_content | String | The body content of the news article | Body |
| **Historical Stock** | | | | |
| 1. | date | Date | The date of the closing price | Year/Month/Day |
| 2. | close | Integer | The closing price of the stock for the day | US Dollars |
| 3. | volume | Integer | The volume of stock traded in the day | Stocks |
| **CPI** | | | | |
| 1. | year | Integer | The year the CPI is recorded | Years |
| 2. | period | String | The order of the month in the year | Months |
| 3. | value | Float | The CPI value | Percentage |

# Potential Feature Engineering & Machine Learning Pipeline

For news, our team chose global topics on war and natural disasters as these topics tend to impact finance industries worldwide. The topic of Jerome Powell is also chosen as stock prices react strongly to his speeches depending on the sentiment. Narrowing down to the company level, we focused on the tech industry and chose META, AAPL and TSLA as these top performing tech stocks tend to be a good representation of the performance of the industry. Sentiment analysis will be performed on the news data to extract useful features. The positive or negative sentiment will then help us predict an upward or downward trend for each stock. We will explore using VADER4, a rule based sentiment analyzer, as well as FinBERT3, a pre-trained model for financial related sentiment analysis based.

Our project’s first model is the Long Short-Term Memory (LSTM), a type of recurrent neural network suitable for time series data. Data on the daily closing price and daily volume traded will be pulled from the yfinance module using Python. We chose closing price over other price indicators as it is the most relevant indicator to determine whether to long or short the stock while daily volume was chosen for additional insight into the prevalence of investor sentiment. This closing price data will be manipulated to generate day-on-day percentage returns. This is then fed into our LSTM which will be trained to predict next-day returns on our selected stocks. Our model will be evaluated by comparing R2 value. This project also aims to explore classical time series models like ARIMA which would allow us to incorporate macroeconomic indicators into the model, since stocks often react to large shifts in these indicators.

We will evaluate our model against a model that was trained without unstructured data with a simulated global war shock. If our model is able to perform better (R-squared valued), this will determine that our model is more resilient to unexpected global events.

1. **Hypotheses**

We have formed the following hypotheses:

1. Integrating text-based models trained with global, industry and company-related news data to existing price prediction models will increase their performance
2. Financial news has a stronger effect on the short-term prices of stocks as compared to long-term prices.
3. Stock returns have a higher level of volatility after the release of sensitive company and industry related news.

# **References**

1. Li K(Y. Predicting stock prices using machine learning. neptune.ai. https://neptune.ai/blog/predicting-stock-prices-using-machine-learning. Published July 22, 2022.
2. Ali W, Wilson J, Husnain M. Micro-, meso- and macro-level determinants of stock price crash risk: A systematic survey of literature. Managerial Finance. https://www.emerald.com/insight/content/doi/10.1108/MF-12-2021-0603/full/html. Published March 15, 2022.
3. Araci, D. (2019, August 27). *Finbert: Financial sentiment analysis with pre-trained language models*. arXiv.org. https://arxiv.org/abs/1908.10063
4. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550
5. Nti IK, Adekoya AF, Weyori BA. A systematic review of fundamental and technical analysis of stock market predictions - artificial intelligence review. SpringerLink. https://link.springer.com/article/10.1007/s10462-019-09754-z. Published August 20, 2019. Accessed October 4, 2022.
6. Lee H, Surdeanu M, MacCartney B, Jurafsky D. On the Importance of Text Analysis for Stock Price Prediction. Retrieved from https://nlp.stanford.edu/pubs/lrec2014-stock.pdf .