Forex Prediction Through News Article Sentiment Analysis and Economic Indicators

Team: Amy Fang, Annie Wang, Ian Lee, Daniel Da **Project Mentor TA:** Jiahao Huang

1) Abstract

This project addresses the complex challenge of forecasting foreign exchange rates between the US Dollar and the Chinese Yuan by integrating diverse and multivariate datasets using Long Short-Term Memory (LSTM) neural networks. Unlike conventional models that primarily rely on historical data, this approach combines four inputs: historical foreign exchange rates, sentiment analysis from Wall Street Journal articles, Google Trends data, and key economic indicators from the US and China. The foreign exchange market, or forex, is the largest and most liquid financial market in the world, where currencies are traded around the clock, enabling companies, governments, and financial institutions to hedge against currency risks, facilitate international trade, and capitalize on economic disparities between countries. Characterized by high volatility, this market reacts dynamically to geopolitical events, economic data, and market sentiment, making it a complex yet crucial component of the global financial system. Our primary target contribution is to collect natural language data from Wall Street Journal, Reddit, and Google Trends to analyze market sentiment. Secondly, we would like to use this data in our LSTM model to forecast the exchange rate, which can eventually be applied to real time data to predict and trade on future exchange rates.

We have compiled a comprehensive dataset of natural language sourced from Wall Street Journal and Reddit from 2004-01-01 to 2024-01-01. With this data, we conducted sentiment analysis to extract the polarities scores of a subset of articles and posts related to Sino-American relations. Additionally, we have gathered Google Trends data for top searches related to the broader economy and foreign exchange markets in both English and Chinese.

2) Introduction

The prediction of foreign exchange rates between two major economies, such as the US and China, is crucial as the volatility and interconnectedness of these markets have implications for international trade, investment decisions, and economic stability. While there is ongoing research centered around making foreign exchange rate predictions using deep learning, conventional approaches typically rely on time-series historical data or sentiment analysis of financial news articles in isolation. The problem is that exchange rate movements are influenced by many interconnected variables at the same time, including economic indicators, geopolitical events, and market sentiments.

This project seeks to address this problem by leveraging Long Short-Term Memory (LSTM) neural networks that encompass a holistic framework incorporating four types of input data: historical exchange rates, news articles from The Wall Street Journal, Google Trends data on terms related to purchasing power parity(PPP), and a range of key economic indicators such as inflation, GDP, unemployment, and public debt in both US and China.

Achieving positive results from taking holistic factors into consideration including sentiment analysis, economic indicators, etc. in forex trading could significantly impact both individual traders and the broader financial market. Enhanced trading strategies that incorporate sentiment data might yield higher returns and improve market timing, thus increasing market efficiency by reflecting a broader range of information in currency prices. This approach could also make markets more resilient to shocks and less prone to rapid, sentiment-driven fluctuations.

Model Inputs	**1) Historical Exchange Rate:** daily exchange rates from 1/1/2000 to 1/1/2024 **2) Natural Language data from News Articles and Social Media:** WSJ articles and Reddit posts/comments related to foreign exchange rates **3) Google Trends data:** normalized number of searches on terms such as " in both the US and China (words translated to Chinese)

	4) Economic Indicators: daily data on economic indicators (inflation, public debt, GDP, unemployment) from both the US and China
Model Outputs	Real foreign exchange rate between US Dollar and Chinese Yuan

3) Background

Machine learning methods have been widely used for predicting financial indicators. Previous works yield significant results using various methods and datasets. Rojas and Herman forecasted the exchange rate between US Dollar and Mexican Peso using market features (such as U.S. bond yields, S&P 500 Index, Dollar Index, etc.) and fundamental variables (such as Consumer Price Index, National Debt, M2 Money Supply, etc.)[Rojas, Herman, 2018][1]. They compared the performance of separate frameworks using these two sets of features respectively and employed multiple models including logistic regression, Support Vector Machines/Regression, Gradient Boosting Classifier/Regression, and Neural Networks. The results showed that Ridge regression and SVM generated the best overall performance, and market features outperformed fundamental variables as predictors. Moreover, the paper points out the direction for improvement as incorporating LSTMs and potentially more variables covering a larger scope of the market.

Meanwhile, some previous works employed sentiment analysis for FX forecasting. Olaiyapo assessed sentiment in social media posts and news articles pertaining to the United States Dollar (USD) using a combination of methods: lexicon-based analysis and the Naive Bayes machine learning algorithm [Olaiyapo, 2024][2]. The findings indicate that sentiment analysis proves valuable in forecasting market movements and devising trading signals. This paper serves as our main inspiration and starting point of the project on exploring sentiment analysis and foreign exchange rates. In addition, Masuda and Takeda focus on predicting the exchange rate using search popularity [Masuda, Takeda, 2019][3]. This paper adopts search terms that are positively or negatively correlated with exchange rates to perform exchange rate forecasting using data collected from Google Trends. We are utilizing some of the analysis on the most important search terms in both the US and China in our project.

Although both the financial indicators and sentiment analysis generated great results, there are few attempts that use them together as predictors. Building on these prior works and following their directions for future improvements, we propose a new model using LSTM while combining both the market data and sentiment analysis from news and social media. To test the effectiveness of the model in response to the changing economic situation, we choose to forecast the exchange rate between US dollars and Chinese Yuan, as there have been significant fluctuations between 2000 to 2024 influenced by all sectors including politics, economy, trade, pandemic, etc.

4) Summary of Our Contributions

Our two main contributions are (1) scraping, collecting, and preprocessing data that encompasses four major areas that may impact exchange rates (2) creating a LSTM model that is able to take into account data from the four major areas we collected data from and predict future foreign exchange rates

1) Data Contribution

Our work involves significant data collection, scraping, and preprocessing. For sentiment analysis, our scraped data is from the Wall Street Journal and Reddit as we wanted data from both a major financial news outlet and a social media platform with finance communities. For popularity from Google search, we collected data using Google Trends on search terms in both languages. For historical data and economic indicators data, we collected from major government or departmental sites.

[1] https://cs229.stanford.edu/proj2018/report/76.pdf
[2] https://doi.org/10.48550/arXiv.2403.00785
[3] https://isf.forecasters.org/wp-content/uploads/gravity_forms/2-dd30f7ae09136fa695c552259bdb3f99/2019/06/Exchange-Rate.pdf

2) Application Contribution

We applied LSTM to a new domain encompassing more areas than traditional approaches by incorporating historical exchange rates, sentiment analysis, economic indicators, and Google search popularity on terms critical for impacting exchange rates. Our final model is a LSTM model that utilizes the data to predict forex rates between the US Dollar and Chinese Yuan.

5) Detailed Description of Contributions

5.1 Methods

We began by collecting data from various sources to use as signals in our model, consisting of natural language data from scraping Wall Street Journal and Reddit, Google Trends, and other relevant economic indicators available online.

Data Scraping and Analysis - Wall Street Journal

To gather data from Wall Street Journal, we first scraped the archive page of Wall Street Journal to collect all article headlines, dates, themes, and URLs. The scraper loops through all dates from 2000-01-01 to 2024-01-01 and sends a request to Wall Street Journal using the Python requests package. With the request response, we use BeautifulSoup to parse the HTML and extract all relevant information for each article. We initially scrape all articles in this timeframe, since some relevant articles on current events may be under categories other than foreign exchange. An example of a row of wsj_headlines.csv is included in [Appendix A].

To gather the actual text data, we ran into many difficulties. Firstly, Wall Street Journal has an extremely robust anti-robot detection system. After many attempts to configure Selenium, we could not find a solution to bypass the anti-robot security on WSJ, since it would prompt the CAPTCHA immediately upon loading the website. Some attempted methods include preloading user cookies, using experimental Chrome options to disable automation detection, running in non-headless mode, and manually solving the CAPTCHAs. Unfortunately, there was no way to access the article through Selenium.

As an alternative, we examined the network activity of the WSJ page to extract the API request that loads the article content. After finding the exact GET request for the text, we ran a new scraper that uses an existing cookie to gather the preview content (that is not behind the WSJ paywall) from each article. Although not ideal, the article content we gathered was substantial enough to perform sentiment analysis on—especially considering that the first few sentences tend to summarize the rest of the article. An example of a row from wsj_fx_articles.csv is included in [Appendix B].

We then separated the Wall Street Journal into two categories: (1) directly related to foreign exchange and (2) related to broader global events and foreign affairs. The first category includes only articles with the theme of "foreign exchange." The latter category includes a broader range of themes, including "markets," "Asia," and "politics." We then filtered all articles to only include those with terms related to China, such as "Yuan," "Renminbi," and "Beijing." We hypothesized that articles with the foreign exchange theme could potentially lag behind the actual exchange rates because their purpose is to directly report on changes in exchange rates. Conversely, broader sentiment on world events would serve as the precursor to potential exchange rate fluctuations.

WSJ Sentiment Analysis - FinBERT

For the sentiment analysis of the Wall Street Journal data, we used FinBERT, which is a pre-trained NLP model to analyze sentiment of financial text. It is a trained BERT language model fine-tuned for financial sentiment classification using a large financial corpus. FinBERT outperforms other models in many financial-related tasks, making it the top choice for sentiment analysis on data from Wall Street Journal[4].

With the three sentiment probabilities, negative, neutral, and positive, we aggregated the data by day and used a moving exponential average to impute data. In particular, articles with the foreign exchange theme were relatively sparse, so imputing data was an extremely important step. Visualizations of sentiment over time for both categories of Wall Street Journal articles

[4] https://www.ijcai.org/proceedings/2020/0622.pdf

are found in [Appendix C] and [Appendix D].

Reddit Sentiment Analysis - VADER

In addition, we sourced raw post and comment data from Reddit[5], including various relevant subreddits to analyze different facets of sentiment. Reddit posts offer a more direct and informal representation of sentiment because it is user-generated and unfiltered, which may reflect public sentiment rather than a more professional perspective. For related content, we used posts and comments from subreddits r/China, r/Forex, and r/News.

Due to the volume of the Reddit data, we opted to use VADER (Valence Aware Dictionary and sEntiment Reasoner) for the sentiment analysis of the text data. VADER is a pre-built lexicon that is tailored to analyze sentiments in social media texts, which makes its application particularly useful for our Reddit data. Additionally, it runs much faster than FinBERT, allowing us to process the large dataset in a reasonable amount of time. Similarly, we noticed that some subreddits were missing data for certain days, particularly in the earlier years. To resolve this issue, we used a moving exponential average to help impute and smooth the data. To see Reddit sentiment over time for r/China, see [Appendix E].

Google Trends Data

Moreover, in an attempt to capture actual market response over time distinct from government-published data sources, we used Google Trends data as a proxy for internet search query volume data. The process of obtaining the necessary data involves several steps. The longest time span where daily data could be obtained from Google Trends is 90 days, and only monthly data for extended time frames (2004 to 2024) is accessible. To overcome this limitation and reconstruct daily-level data over this period, we utilized the Google Trends API, specifically the **'pytrends'** library to acquire both monthly level data over the extended period and daily-level over intervals of 90 days. Subsequently, we utilized overlapping periods to deduce normalized daily data[6]. To select the search terms relevant to exchange rates, we referenced Bulut's research[Bulut 2017] on Google Trends and the forecast of exchange rate models[7] and selected search terms to capture the Purchasing Power Parity(PPP), money demand and money supply.

Price-related	Inflation, Prices, CPI, Cheap
Income-related	Buy, Spend, Save, Donate, Job, Vacation, Foreclosure
Liquidity-related	Cash, Credit, ATM

These terms are further translated into Chinese to capture the market response within non-English search terms.

Economic Indicator Data

This data include major economic indicators published by the U.S. and China. To capture the national economy and growth, five major indicators are used for both countries: interest rates, GDP, unemployment rate, inflation rate, and public debt. The data is collected from the following sources:

U.S. Bureau of Economic Analysis (BEA)
The Federal Reserve Bank of St. Louis (FRED)
U.S. Bureau of Labor Statistics (BLS).
U.S. Department of the Treasury
The National Bureau of Statistics of China
Trading Economics

[5] https://the-eye.eu/redarcs/
[6] https://towardsdatascience.com/reconstruct-google-trends-daily-data-for-extended-period-75b6ca1d 3420
[7] Bulut L. Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting*. 2018; 37: 303–315. https://doi.org/10.1002/for.2500

These indicators help to assess the overall economic performance, price stability, employment situation, investment climate, and fiscal health of a nation. The collection of these data provides a comprehensive view into the drivers of and pressures on national economic growth and living standards over time.

5.2 Experiments and Results

The model we chose to model the exchange rate is LSTM, and there are several reasons it fits well with our goal: 1) LSTM networks are designed to handle sequential data and can effectively learn and model long-range dependencies and patterns in time series data like exchange rates. This allows them to better capture the complex temporal dynamics present in currency markets. 2) LSTM models can learn and approximate highly non-linear functions, which is crucial for accurately forecasting the non-linear and chaotic behavior often exhibited by exchange rates. 3) LSTMs can incorporate multiple input features like technical indicators, macroeconomic factors, news sentiment etc. to improve forecasting accuracy by leveraging different data sources. 4) Robust to noise: The gating mechanism in LSTMs allows them to be robust to noisy and irrelevant data inputs, which is beneficial when dealing with volatile financial time series.

By leveraging the ability to model long-range patterns, handle non-linearity, and integrate multivariate inputs, LSTM serves as a good model for forecasting the exchange rate using the data we gathered. For a more detailed diagram of the LSTM layers used, see [Appendix F].

Our hypothesis is that the model combining four categories of inputs (Historical Exchange Rate, Natural Language data from News Articles, Google Trends data, and Economic Indicators) will result in more accurate predictions than what the baseline model gives. We used LSTM for both frameworks but used different datasets to make comparisons. The baseline model takes into account historical exchange rates and economic indicators, whereas the new model also takes into account Google Trends data and sentiment analysis of Natural Language data. The performance metrics we use for model evaluation are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R2 Score. The combination of these metrics gives a comprehensive evaluation of the models as each captures different aspects of model performance. MSE shows how well the model fits the data on average, whereas RMSE provides an easier interpretation than MSE since it represents the typical magnitude of the error. MAE calculates the average absolute difference between predicted and actual values, and is less sensitive to outliers compared to MSE/RMSE. R2 Score measures the goodness-of-fit of the given model. An R2 of 1 indicates the model perfectly fits the data, while 0 means the model is no better than using the mean of the target variable.

We note that LSTM and neural networks in general are prone to overfitting. Since there is the potential of a large amount of noise in all of our data, we utilize two techniques to combat overfitting. We add a dropout layer in between our LSTM layers, which randomly sets a fraction of the input weights to 0. We also add an early stopping mechanism on our validation data.

Our full dataset of signals begins 2004-01-01 and ends 2024-01-01. We split our data into (1) training, which includes data from 2004-01-01 to 2019-12-31, (2) validation, which includes data from 2020-01-01 to 2021-12-31, and (3) test, which includes data from 2022-01-01 to 2023-12-31.

Below is a table summarizing the results of both the baseline and complete model:

	MSE	RMSE	MAE	R2
Sentiment Model - 7 day timeframe	0.0127	0.1131	0.09017	0.2240
Baseline Model - 7 day timeframe	0.0441	0.2100	0.1668	-1.6720
Sentiment Model - 30 day timeframe	0.0077	0.0880	0.0717	0.2431

Baseline Model - 30 day timeframe	0.01750	0.1322	0.1147	-0.7081
Sentiment Model - 90 day timeframe	0.0101	0.1006	0.0810	0.0104
Baseline Model - 90 day timeframe	0.01444	0.1201	0.0934	-0.4098

The results we get supports our hypothesis: the proposed model with the dataset of 4 categories outperforms the baseline model that doesn't implement sentiment analysis. This also confirms the expectations in the works we cited earlier, that incorporating more diverse data can lead to better predictions as they capture a wider scope of the market situation. The sentiment analysis plays a key role in capturing the nuances in all sectors, including the economics, foreign policies, trading, business, etc. that are otherwise hard to quantify and evaluate. To see a comparison of the predicted rates to the actual rates, see [Appendix G] and [Appendix H].

6) Compute/Other Resources Used

We were able to run our model using GPUs on Google Collab. It did take a while to scrape WSJ article data and run our model so for similar models with bigger datasets, greater memory and computational power might be required. For sentiment analysis, we used Google Collab Pro to get access to more RAM.

7) Conclusions

Our complete model that was trained on 4 different types of data had better performance than the baseline model that was trained on all the data besides Natural Language Data. Our model reveals the benefits of including sentiment analysis into forex predictions.

There are many ways to expand and improve on our model in the future. The model can be trained to predict rates between different countries. There is a lot of Reddit and WSJ data on China and it would be important to see if these sources of data are also helpful when predicting forex rates between the US and another country. Another way to expand on our project would be to see if these data sources can be used to predict other financial instruments such as stocks, commodities, and bonds.

Our model relies on data that we have scraped or collected from various sources, and data privacy is a key ethical consideration. It is important to maintain complete anonymity when scraping data from Reddit and to protect user's privacy. Scraping articles from Wall Street Journal also poses an ethical dilemma since their terms of service prohibits automated scraping of their content: scraping WSJ content without their approval infringes on intellectual property rights. Any similar models in the future would have to make sure to be transparent about where the data came from and how it was used, and to be compliant with all privacy regulations.

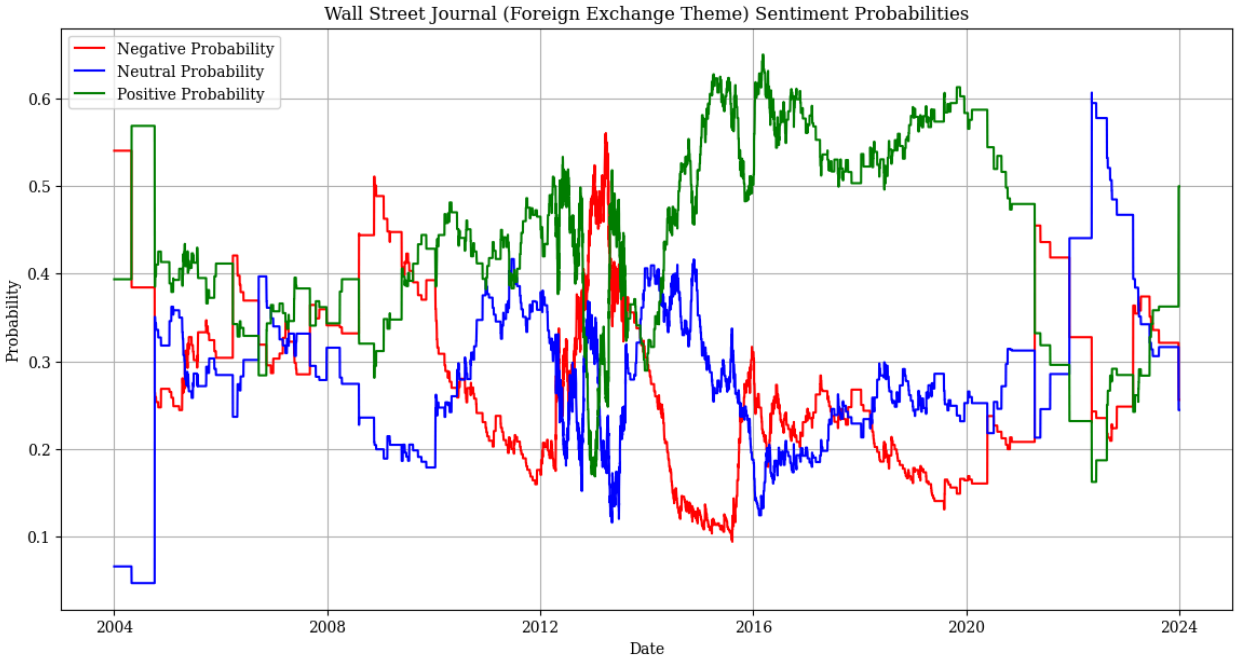
Being able to better predict forex rates results in a more stable global financial system, benefiting businesses and global economies. More precise forex predictions reduce information asymmetry and overall volatility of currencies. Businesses will be able to manage currency risk better, enabling them to increase investments. Central banks can also pass more beneficial monetary policy since forex rates are connected to interest rates, inflation, and overall economic growth.

Date Published	Headline	URL	Theme	Time Published
2000-01-04	"Bank of Japan Moves To Curb Dollar's Fall"	https://www.wsj.com/articles/SB946912050945363247	Foreign Exchange	5:58 AM ET

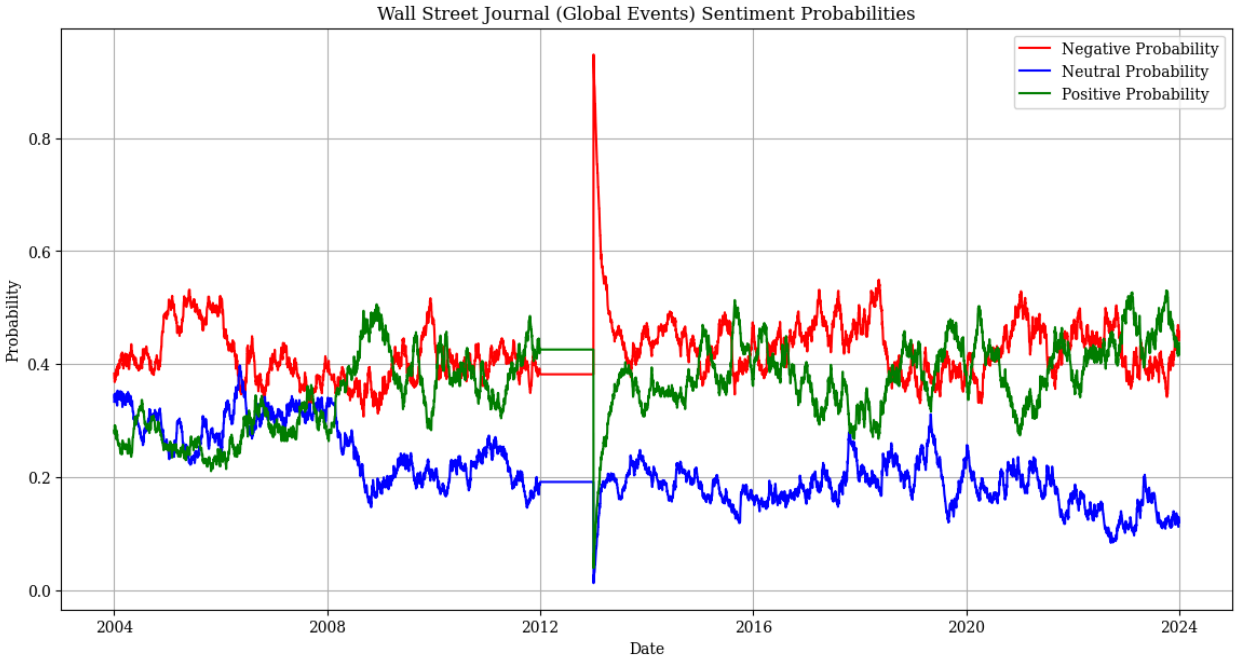
Appendix A: Example of Wall Street Journal headline data

Date Published	Headline	Theme	Article Content
2018-02-07	"US dollar rises following budget agreement"	Foreign Exchange	"The U.S. dollar rose after Senate leaders announced a two-year budget agreement that pushed back concerns that a partisan stalemate could lead to a government shutdown or a debt default. The Wall Street Journal Dollar Index, which measures the currency against a basket of 16 others, posted its third gain in four days, rising 0.5% to 84.23. Even as the dollar gained, rising 0.9% against the euro, it declined 0.2% against the Japanese yen, as investors continued to seek haven assets after a recent surge in volatility roiled financial markets."

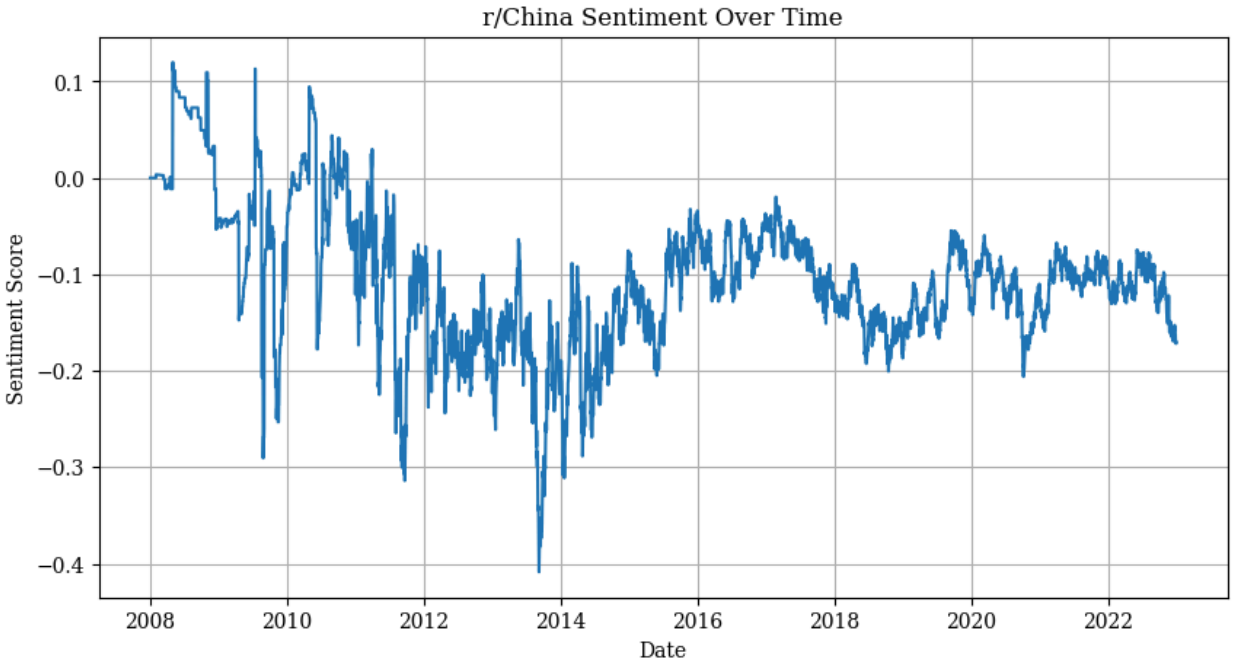
Appendix B: Example of Wall Street Journal article data



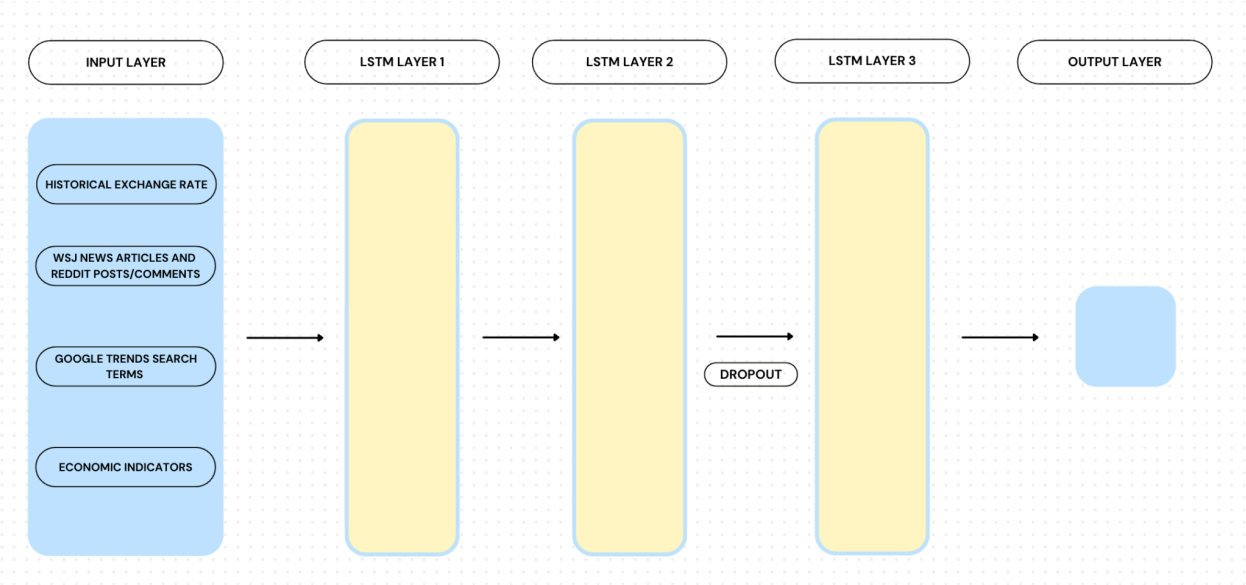
Appendix C: Polarity probabilities over time of forex-related Wall Street Journal articles.



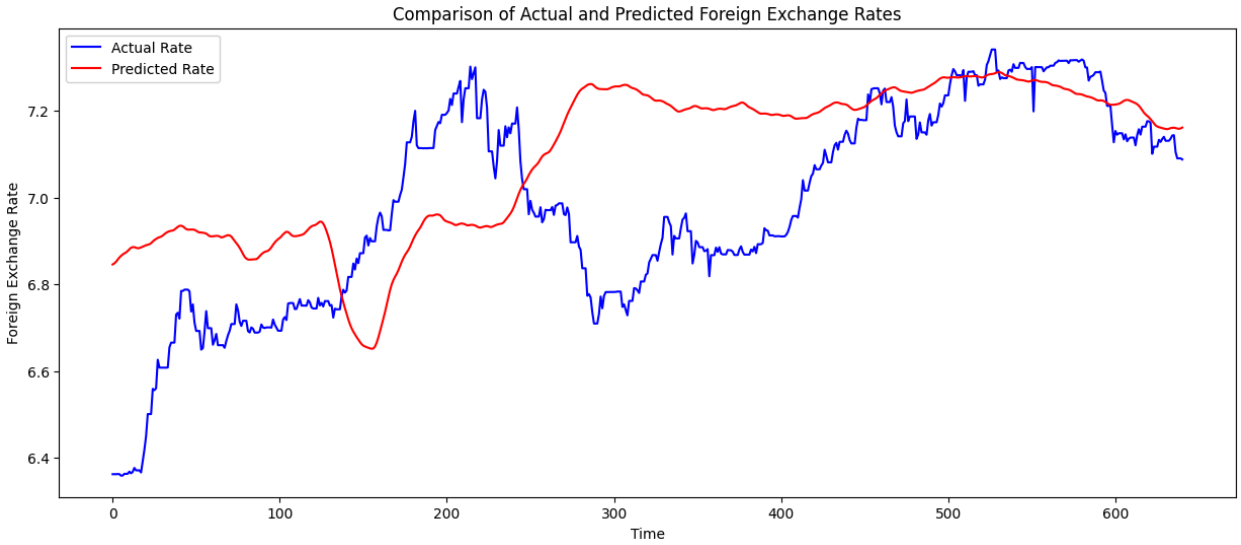
Appendix D: Polarity probabilities over time of global events articles in Wall Street Journal.



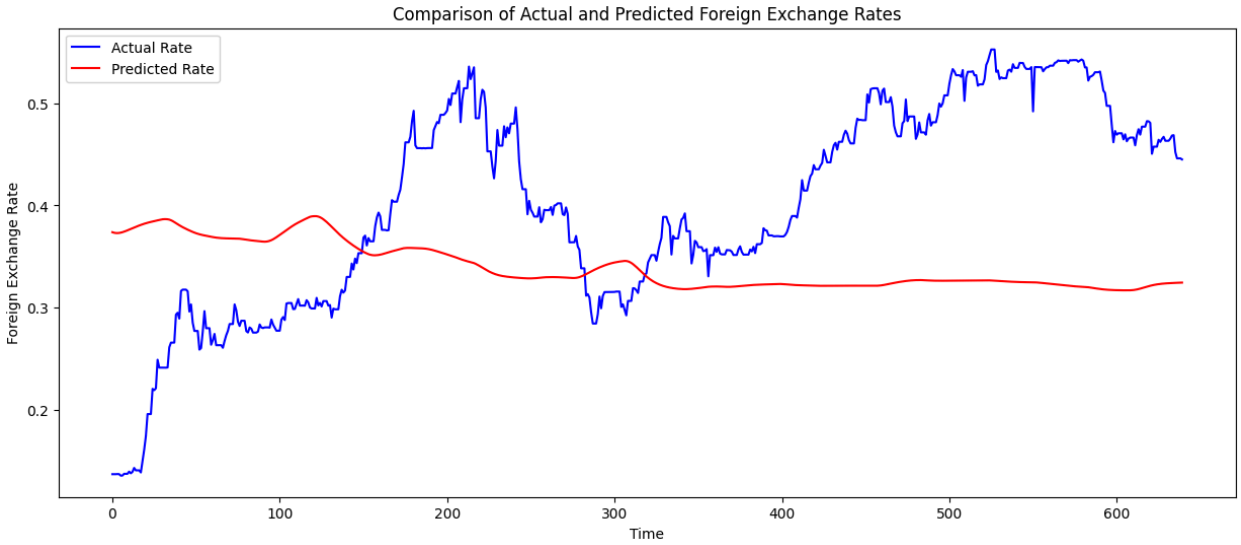
Appendix E: Reddit sentiment over time in subreddit r/China.



Appendix F: Diagram of LSTM model.



Appendix G: Performance of the sentiment model for a timestep of 30 days on the test data beginning from 2022-01-01.



Appendix H: Performance of the baseline model for a timestep of 30 days on the test data beginning from 2022-01-01. The model is unable to capture the trends of the exchange rate.

Other Prior Work

1. Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.
2. Joshi Kalyani, Prof. H. N. Bharathi, Prof. Rao Jyothi, Stock trend prediction using news sentiment analysis. DOI: https://doi.org/10.48550/arXiv.1607.01958
3. Seifollahi, S., Shajari, M. Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction. *J Intell Inf Syst* 52, 57–83 (2019). https://doi.org/10.1007/s10844-018-0504-9
4. Foreign Exchange Forecasting via Machine Learning https://cs229.stanford.edu/proj2018/report/76.pdf
5. "Impact of News Sentiment on Foreign Exchange Rate Prediction" by A. Tadphale, H. Saraswat, O. Sonawane and P. R. Deshmukh
6. https://ieeexplore.ieee.org/document/10205534
7. Bulut L. Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting*. 2018; 37: 303–315. https://doi.org/10.1002/for.2500
8. T. Sidogi, R. Mbuvha and T. Marwala, "Stock Price Prediction Using Sentiment Analysis," *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Melbourne, Australia, 2021, pp. 46-51, doi: 10.1109/SMC52423.2021.9659283.
9. Deng, X., Bashlovkina, V., Han, F., Baumgartner, S., & Bendersky, M. (2023). What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis. *Companion Proceedings of the ACM Web Conference 2023*, 107–110. https://doi.org/10.1145/3543873.3587324

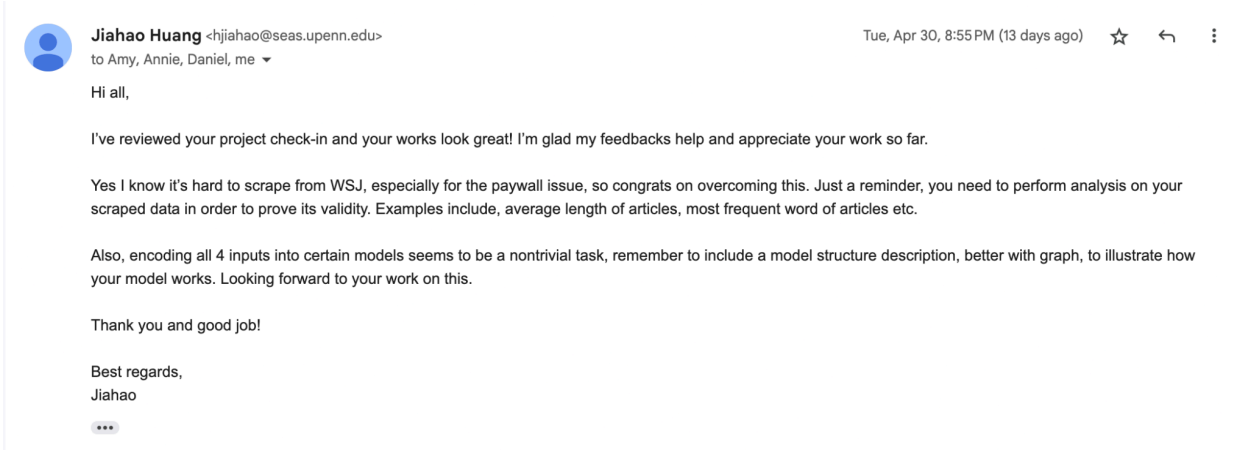
Broader Dissemination Information:

Your report title and the list of team members will be published on the class website. Would you also like your pdf report to be published?
YES

(Exempted from page limit) **Work Report: This may look like your GANTT chart from the midway report, with more completed steps now. Okay to modify.** (Mark completed steps in green, as shown here. For convenience, you may split into two charts, one till Nov 8, and another for after Nov 8, placed one below the other.)

PERSON (S)	TASK (S)	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7
		March	April				May	
Daniel, Annie	Scrape sentiment analysis data from Reuters, WSJ, and Bloomberg							
Daniel, Annie	Get and Clean Google Trends Data							
Ian, Amy	Clean and process foreign exchange rate historical data							
Ian, Amy	Locate, clean, and process other market indicators							
All	Baseline model for forex rate predictions							
Whoever's Free	Preprocess Sentiment Analysis Data							
All	Write Check-in							
Daniel, Annie	Finalize which Google Search and Economic Indicator Data will be used							
All	Work on final models							
All	Work on final deliverables							

(Exempted from page limit) Attach your midway report here, as a series of screenshots from Gradescope, starting with a screenshot of your main evaluation tab, and then screenshots of each page, including pdf comments. This is similar to how you were required to attach screenshots of the proposal in your midway report.



Jiahao Huang <hjiahao@seas.upenn.edu> Tue, Apr 30, 8:55 PM (13 days ago)
to Amy, Annie, Daniel, me

Hi all,

I've reviewed your project check-in and your works look great! I'm glad my feedbacks help and appreciate your work so far.

Yes I know it's hard to scrape from WSJ, especially for the paywall issue, so congrats on overcoming this. Just a reminder, you need to perform analysis on your scraped data in order to prove its validity. Examples include, average length of articles, most frequent word of articles etc.

Also, encoding all 4 inputs into certain models seems to be a nontrivial task, remember to include a model structure description, better with graph, to illustrate how your model works. Looking forward to your work on this.

Thank you and good job!

Best regards,
Jiahao

Forex Prediction Through Social Media and News Article Sentiment Analysis

Team: Amy Fang, Annie Wang, Ian Lee, Daniel Da **Project Mentor TA:** Jiahao Huang

1) Introduction

Using sentiment analysis on the text data gathered from the Internet, we aim to predict the foreign currency exchange rate between USD and JPY.

Model Inputs	**1) Historical Exchange Rate:** daily exchange rates from 1/1/2000 to 1/1/2024 **2) Natural Language data from News Articles:** news article headlines and social media data from Twitter and Reddit **3) Google Trends data:** normalized number of searches on terms such as " in both the US and Japan (words translated to Japanese) **4) Economic Indicators:** daily data on economic indicators including interest rates, GDP, unemployment rates from both countries.
Model Outputs	Prediction of future exchange rate

To train our data, we will use historical data from January 1st, 2000 to January 1st, 2020. Our testing data will be data from January 1st, 2020 to January 1st, 2024. Notably, there are a wide range of macroeconomic trends and current events in our testing time period (COVID-19, Ukraine-Russia conflict, China's economic slowdown, etc.) which can test if our model is able to accurately capture the impact of these deviations on foreign exchange trading.

To evaluate our model's performance, we can use two approaches: backtesting or forward testing. For feasibility purposes, we will focus on backtesting our model with the data from 2020-2024. The primary metric we will use to evaluate performance will be deviation of predicted values from actual values such as MAPE and RMSE.

Motivation: Achieving positive results from taking holistic factors into consideration including sentiment analysis, economic indicators, etc. in forex trading could significantly impact both individual traders and the broader financial market. Enhanced trading strategies that incorporate sentiment data might yield higher returns and improve market timing, thus increasing market efficiency by reflecting a broader range of information in currency prices. This approach could also make markets more resilient to shocks and less prone to rapid, sentiment-driven fluctuations.

2) How We Have Addressed Feedback From the Proposal Evaluations

Our feedback was organized into three bullet points:

- The model inputs and outputs are unclear

    - We have clarified our input and output variables, as seen in the table above. We decided to predict the exact value of the exchange rate and not just a positive or negative signal.
- The data collection process will be very difficult
    - We were able to successfully scrape a large amount of data from the Wall Street Journal. While the data collection process is definitely difficult, we also hope to gather some more data from a social media platform (Twitter or Reddit) for another dimension of market sentiment.
- RNN Comparison is not fair or valid
    - We've re-evaluated and modified our second contribution. Rather than comparing RNN to our baseline model of using just rates, we will instead be applying neural network models to a new domain encompassing more areas than traditional approaches by incorporating historical exchange rates, sentiment analysis, economic indicators, and Google search popularity. We are committing to come up with a valid neural network solution that will accurately be able to predict forex. We will measure the performance of our model by calculating RMSE and MAPE of predicted values from actual values and hope to minimize these accuracy errors.

3) Prior Work We are Closely Building From

A. Olaiyapo O. F. Applying news and media sentiment analysis for generating Forex trading signals. *Review of Business and Economics Studies*. 2023;11(4):84-94. DOI: 10.26794/2308-944X-2023-11-4-84-94
    a. This paper explores how news and media sentiment analysis can be used to predict forex trading signals. It serves as our main inspiration and starting point of the project on exploring sentiment analysis and foreign exchange rates.

B. "Application of Google Trends Data in Exchange Rate Prediction" by Motoki Masuda and Fumiko Takeda: https://isf.forecasters.org/wp-content/uploads/gravity_forms/2-dd30f7ae09136fa695c552259bdb3f99/2019/06/Exchange-Rate.pdf
    a. This paper discusses applying search popularity on search terms positively or negatively correlated with exchange rates to perform exchange rate forecasting using data collected from Google Trends. We are utilizing some of the analysis on the most important search terms in both the US and Japan in our project.

4) What We are Contributing

Past research in exchange rates predictions often involves traditional approaches of solely using time series data, or sentiment analysis. In our project, we make an attempt to explore exchange rate prediction by encompassing more domain areas including:

- sentiment analysis from news and social media
- popularity from Google search on terms impacting exchange rates
- historical exchange rate data
- economic indicators

Our two main contributions are (1) collecting data for sentiment analysis and (2) applying neural net models into a new combined domain encompassing more areas than traditional approaches.

1) Data Contribution

Our work involves significant data collection, scraping, and preprocessing. For sentiment analysis, our data scraped will primarily come from major financial news outlets, such as Wall Street Journal, and social media platforms with finance communities, such as Twitter or Reddit. For popularity from Google search, we collected data using Google Trends on search terms in both languages. For historical data and economic indicators data, we collected from major government or departmental sites.

2) Application Contribution

We will be applying neural network models such as LSTM, Transformers to a new domain encompassing more areas than traditional approaches by incorporating historical exchange rates, sentiment analysis, economic indicators, and Google search popularity on terms critical for impacting exchange rates. Our final model will be one of these models or a combination of them that best predicts future exchange rates.

5) Detailed Description of Each Proposed Contribution, Progress Towards It, and Any Difficulties

Data Contribution:

In the data collection process, we have focused on gathering relevant articles from Wall Street Journal. First, we scraped the archive page of Wall Street Journal to collect all article headlines, dates, themes, and URLs. The scraper loops through all dates from 2000-01-01 to 2024-01-01 and sends a request to Wall Street Journal using the Python requests package. With the request response, we use BeautifulSoup to parse the HTML and extract all relevant information for each article. We initially scrape all articles in this timeframe, since some relevant articles on current events may be under categories other than foreign exchange.

Below is an example of a row of our wsj_headlines.csv.

Date Published	Headline	URL	Theme	Time Published
2000-01-04	"Bank of Japan Moves To Curb Dollar's Fall"	https://www.wsj.com/articles/SB946912050945363247	Foreign Exchange	5:58 AM ET

Next, we performed some preliminary data cleaning to prepare to gather the article content of each article. For now, we only include articles with the Foreign Exchange theme to test out our actual text scraper by pre-processing the dataset to only include rows where the headlines contain words such as “FX,” “foreign exchange,” “forex,” etc.

To gather the actual text data, we ran into many difficulties. Firstly, Wall Street Journal has an extremely robust anti-robot detection system. After many attempts to configure Selenium, we could not find a solution to bypass the anti-robot security on WSJ, since it would prompt the CAPTCHA immediately upon loading the website. Some attempted methods include preloading user cookies, using experimental Chrome options to disable automation detection, running in non-headless mode, and manually solving the CAPTCHAs. Unfortunately, there was no way to access the article through Selenium.

As an alternative, we examined the network activity of the WSJ page to extract the API request that loads the article content. After finding the exact GET request for the text, we ran a new scraper that uses an existing cookie to gather the content (that is not behind the WSJ paywall) from each article.
Below is an example of a row from wsj_fx_articles.csv:

Date Published	Headline	Theme	Article Content
2018-02-07	"US dollar rises following budget agreement"	Foreign Exchange	"The U.S. dollar rose after Senate leaders announced a two-year budget agreement that pushed back concerns that a partisan stalemate could lead to a government shutdown or a debt default. The Wall Street Journal Dollar Index, which measures the currency against a basket of 16 others, posted its third gain in four days, rising 0.5% to 84.23. Even as the dollar gained, rising 0.9% against the euro, it declined 0.2% against the Japanese yen, as investors continued to seek haven assets after a recent surge in volatility roiled financial markets."

To collect data on google search popularity on search terms impacting exchange rates, we utilized Google Trends. We collected data on search terms (including: "investment," "crisis," "inflation," "GDP") in the US and translated the terms to collect data on search popularity in Japan.

Application Contribution:
For our application contribution, we have our baseline RNN model that makes predictions based off of just past exchange rate data. Now that we are close to done with our data contribution, we will shift our focus into developing our final model.

Our hypothesis is that the model combining four categories of inputs (Historical Exchange Rate, Natural Language data from News Articles, Google Trends data, and Economic Indicators) will result in more accurate predictions than the baseline model gives, which only takes the past exchange rate as inputs. We will validate this hypothesis by comparing the MAPE between the baseline model and the advanced models.

6) Risk Mitigation Plan

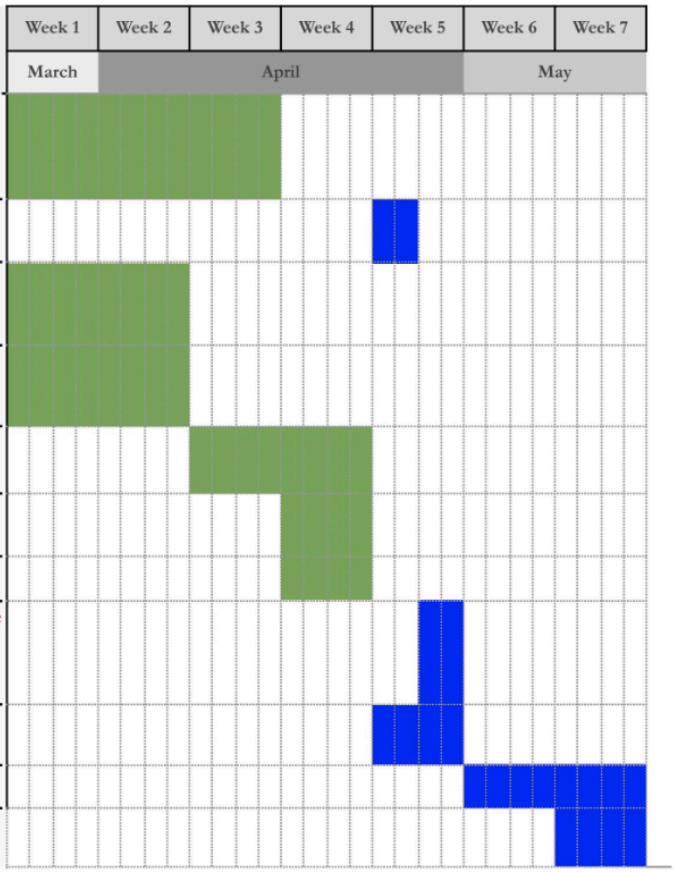
We already have a baseline model factoring just exchange rates in, and have enough news article natural language data to build another model that factors those in. While collecting more macroeconomic data and Google Trends, Tweets and Reddit, we can start training a simplified model with the existing exchange rates and news articles to get some early results. After gathering the full dataset, we will try the baseline and advanced algorithms on a small subset of our data and assess the accuracy. There are a few models we would like to attempt, such as transformers, LSTM, and combined models. If one model does not work, we will pivot to other models. During this stage, we will also try and evaluate the combination of variables that result in the best prediction. After testing out on the subset of data and determining the variables to include, we will expand the training data to the full dataset we gathered. If the amount of computation is not manageable, we will decrease the time span of the exchange rate we are using in the training dataset and therefore decrease the amount of computation needed. Since our exchange rate training dataset only includes the years 2000 - 2020, the amount of data is unlikely to be overwhelming.

(Exempted from page limit) Other Prior Work / References (apart from Sec 3) that are cited in the text:

1. Joshi Kalyani, Prof. H. N. Bharathi, Prof. Rao Jyothi, Stock trend prediction using news sentiment analysis. DOI: https://doi.org/10.48550/arXiv.1607.01958
2. Seifollahi, S., Shajari, M. Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction. *J Intell Inf Syst* 52, 57–83 (2019). https://doi.org/10.1007/s10844-018-0504-9
3. Foreign Exchange Forecasting via Machine Learning
   https://cs229.stanford.edu/proj2018/report/76.pdf
4. "Impact of News Sentiment on Foreign Exchange Rate Prediction" by A. Tadphale, H. Saraswat, O. Sonawane and P. R. Deshmukh
   https://ieeexplore.ieee.org/document/10205534

(Exempted from page limit) **Full Work Plan**:

PERSON (S)	TASK (S)
Daniel, Annie	Scrape sentiment analysis data from Reuters, WSJ, and Bloomberg
Daniel, Annie	Scrape Twitter and Reddit data
Ian, Amy	Clean and process foreign exchange rate historical data
Ian, Amy	Locate, clean, and process other market indicators
All	Baseline model for forex rate predictions
Whoever's Free	Preprocess Sentiment Analysis Data
All	Write Check-in
Daniel, Annie	Finalize which Google Search and Economic Indicator Data will be used
Ian, Amy	Work on simple model
All	Work on final models
All	Work on final deliverables



(Exempted from page limit) Attach your proposal here and feedback TAs gave you.

Forex Prediction Through Social Media and News Article Sentiment Analysis

Amy Fang, Annie Wang, Ian Lee, Daniel Da

Task T:
We propose to predict the foreign currency exchange rate between USD and EUR, GBP, RMB, and JPY, CAD based on the weekly exchange rate and the sentiment analysis of the news articles during the past year.

Experience E:
Our project will be split up into two phases: (1) data collection and processing and (2) algorithm and model development.
For Phase 1, we'll need to scrape and collect data from the past decade for sentiment analysis. Some possible sources include Reuters, Wall Street Journal, Bloomberg, and other public news sources. Other social media sources could include Twitter, StockTwits (Twitter for investors), and Reddit. After collection, we will clean the data (lowercasing and removing punctuation and special characters), tokenize the text, and possibly conduct entity extraction/recognition.
For Phase 2, we'll apply different sentiment analysis techniques onto our data to generate signals for foreign exchange rate predictions. We will reserve the most recent two years of data for testing. We will focus on recurrent neural networks (RNN) for the majority of our project, but may evaluate other models for comparison, such as Naive Bayes, CNNs, or ensemble models. From our sentiment analysis, we would gather and test relevant signals (sentiment polarity, word counts, possibly a custom weighting with VADER) as predictors for our foreign exchange pairs.

Performance metrics P:
We will separate historical data into train and test data and train our model on the train data. We can then look at how our model does on the test data in terms of accuracy (Mean Squared Error) and the difference between training and test data performance.
We will also compare our model to a base model, which will be a recurrent neural network that only takes into account past exchange rate data. Then, we will be able to measure how much better sentiment analysis makes our model compared to the base.

Prior work:

1. Olaiyapo O. F. Applying news and media sentiment analysis for generating Forex trading signals. *Review of Business and Economics Studies.* 2023;11(4):84-94. DOI: 10.26794/2308-944X-2023-11-4-84-94
2. Joshi Kalyani, Prof. H. N. Bharathi, Prof. Rao Jyothi, Stock trend prediction using news sentiment analysis. DOI: https://doi.org/10.48550/arXiv.1607.01958
3. Seifollahi, S., Shajari, M. Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction. *J Intell Inf Syst* 52, 57–83 (2019). https://doi.org/10.1007/s10844-018-0504-9

Nature of main proposed contribution(s):
Data Collection:
We will need to collect a lot of data for sentiment analysis, and the quality of this data will have a huge impact on how our model does. We will need to scrape keywords off of social media platforms such as Reddit and

Twitter along with articles from reputable news companies such as the Wall Street Journal and Financial Times. It will be important to scrape data from similar sources throughout the time period we are looking at.

Application:
We will apply our model to the exchange rate data and sentiment data to predict the foreign exchange market. We hope that the model will be able to predict an increase or decrease from the previous time and be able to predict the amount of that change to be as close to the actual value as possible. Moreover, by constructing and training our model we can separate the data from social media platforms and the news companies and determine how they are weighted differently and the impact that they have on the final prediction.

Why we care:
Ian is interested in predicting possible arbitrage opportunities in the currency exchange market. Amy is passionate about natural language processing. Annie is a finance major and is fascinated about how markets shift to real world events. Daniel loves traveling and wants to get a good deal on vacation. In general, the foreign exchange market and Forex trading is highly influenced by global news events, economic indicators, and market sentiment. By incorporating sentiment analysis into the prediction in addition to the historical data, we can potentially better capture the market dynamics and make more informed decisions.

Which parts of the curriculum from this class do you expect to apply?:
Preprocessing: splitting up the data into train, validation, and test sets for later ML and fine tuning. We foresee the usage of NLP techniques covered in class later when scraping and processing the social media texts.
Model: we plan to use Neural Networks with fine tuning to perform predictive analysis. We would first build the baseline model without considering the sentiment analysis and assess the performance. Then, we would like to add the sentiment analysis into the model and observe the changes in accuracy of the predictions.

Compute Requirements:
We foresee the majority of the work done in Google Colab. A moderately powerful CPU should be enough for our project, but the data volume for sentiment analysis may be large. For what we foresee, our personal machines should suffice.

Expected challenges and risk mitigation:
One of the primary challenges is the unstructured nature of news and media data and the difficulty of cleaning them to obtain useful information. The inherent noise in the data may affect the accuracy of our sentiment analysis and prediction. To mitigate this risk, we will make sure our preprocessing is robust, using techniques to minimize noise and perform feature selection. Additionally, we will evaluate the performance of our model with sentiment analysis on prediction solely using historical data and conduct backtesting to ensure the reliability of the model.

Ethical considerations and broader social impact:
One important ethical consideration is scraping news and media sentiment in concerns about data privacy and content ownership. While news articles and public media most may be considered publicly available information, we must make sure to comply with regulations and ethical considerations.
A potential good social impact is the ability to better forecast exchange rates and evaluate how media sentiments contribute to the exchange rate markets and forecasts.

Feedback:

I've reviewed your proposal and find your project exciting, great work! I want to give some brief feedbacks to make sure your later submission meet the requirements.

1. Your problem statement is generally clear, But for the model part, I hope you could have a more detailed description of the pipeline, e.g. what would be the input and output for your model, I'm a little confused on if you want to predict a positive/negative signal of forex trend or the exact value of the exchange.
2. The problem you're trying to solve is compelling but according to my experience, the data collection stage will be very hard. I suggest you start early if possible and focus on only some of the media in order to come up with a consistent scraping rule. I'm looking forward to your work on this.
3. I find your comparison of RNN with media input and baseline with rate value input less intuitive in terms of fairness and validity. I suggest you could put some thoughts on this issue. Or, coming up with a valid RNN solution is good enough from my perspective.

Overall good job for the proposal! Looking forward to seeing an outstanding course project from you in the end!