## **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

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
   1. 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.
2. “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>
   1. 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.

1. 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](https://arxiv.org/search/cs?searchtype=author&query=Kalyani,+J), [Prof. H. N. Bharathi](https://arxiv.org/search/cs?searchtype=author&query=Bharathi,+P+H+N), [Prof. Rao Jyothi](https://arxiv.org/search/cs?searchtype=author&query=Jyothi,+P+R), 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

1. “Impact of News Sentiment on Foreign Exchange Rate Prediction” by A. Tadphale, H. Saraswat, O. Sonawane and P. R. Deshmukh

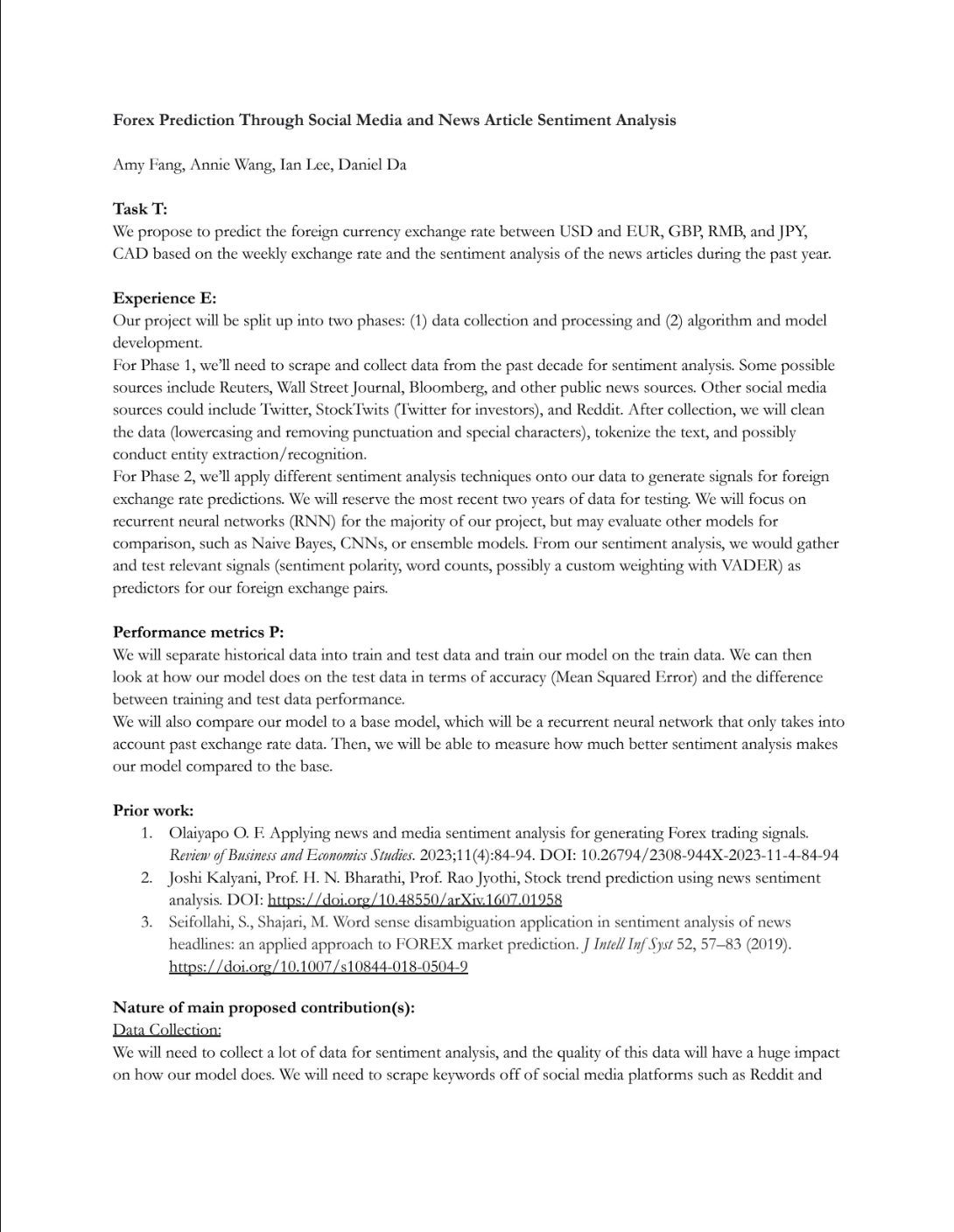
https://ieeexplore.ieee.org/document/10205534

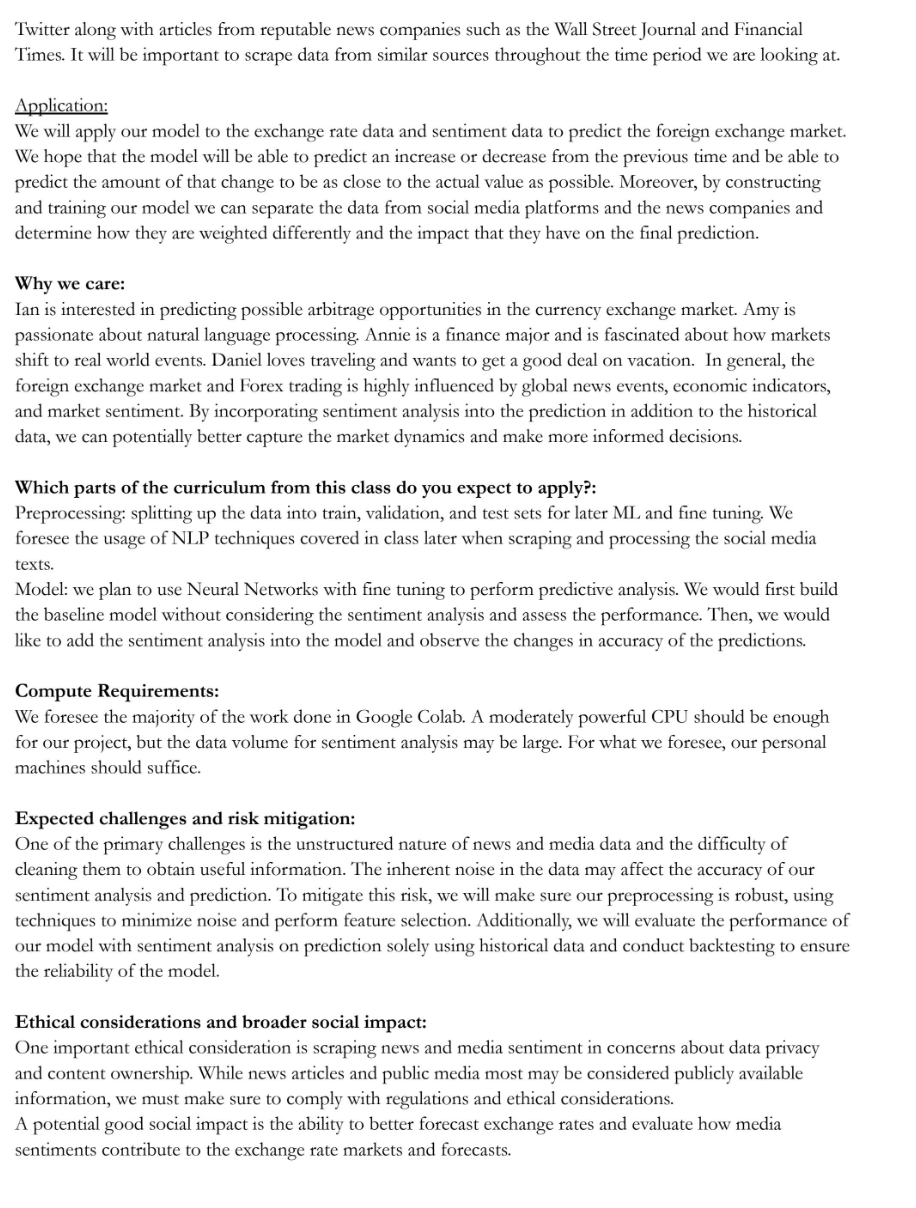
##### 

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

| **PERSON (S)** | **TASK (S)** | **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** | 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.





Feedback:

