NLP Final Report

20221062 온재현, 20221096 이현서

1. Creating Dataset

A. Pre-processing

First, we retrieved the HTML fragments containing the article text and date data from the CSV table. Before proceeding with tokenization, we used the find method to identify and slice out unnecessary data. The unnecessary data often included company information included for formatting purposes (starting with “About f{firm_name}”), hyperlinks or guidance for more information (e.g., “for more information, ...”), contact information of the article writer, and license information. Additionally, regular expressions were used to locate the date data.

B. shape conversion

For all data where date information could be found, text-date pairs were created. The initial dataset was formed as a list of lists with [stock name, text, date] as one data unit. Considering that the data had a chronological order, data before a certain date was used as the training set, and data after that date was used as the validation set.

Next, we directly defined a dataset class that paired texts with stock price movements. Using the date data, we calculated the open-close values from the stock price table retrieved from yfinance to determine whether the stock price went up or down, labeling them as 0 or 1 (binary classification model). To explore more variations, we also attempted a three-class classification model where values above a certain percentage 'a' were labeled as 2, values between a% and -a% were labeled as 1, and values below -a% were labeled as 0.

While verifying the labeling results, we noticed that a significant amount of data was missing. This was because, on non-trading days, the corresponding dates could not be found in the yfinance table. To handle this error, we implemented a method that incremented the date value until a valid date could be read when the date could not be found.

2. Model settings and training

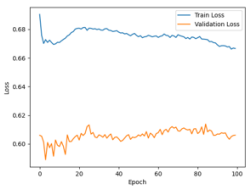
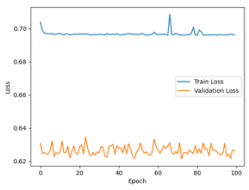
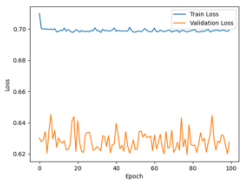
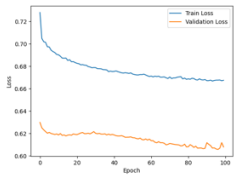
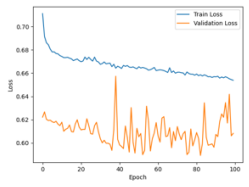
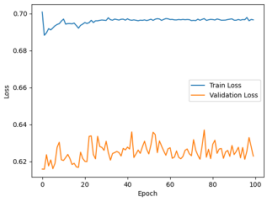
A. Model, Optimizer, Scheduler

Model used: Transformer (BertForSequenceClassification)
Optimizer: Adam, SGD
Scheduler: CosineAnnealing

During several initial attempts, we encountered difficulties where training did not progress at all (loss did not decrease) depending on the lr setting. To mitigate this, we used the CosineAnnealing Scheduler, which is widely used with transformer models, for reasonable lr settings. Considering that this project involves a very high computational load, we also used SGD, which can update more quickly to near the target point, in addition to Adam.

Initially, we used the trainer class supported by Huggingface with adjusted parameters, but we found it difficult to easily track the progress of the training. To resolve this, we directly implemented a training loop consisting of stages such as training step and validation, and added train/valid loss values at each epoch to plot the final training results using plt.

B. Comparison of results according to Lr, Optimizer settings (at three-class classification)

	5e-5	4e-4	3e-3
Adam_			
SGD :			

In our three-class classification model, we found that changing the learning method and learning rate did not reduce the loss below a certain level.

C. Binary Classification Model Results (Examples)

Generated Text	Prediction (0: down, 1: up)
XYZ Corporation Announces Stock Buyback Program New York, NY - June 14, 2024 - XYZ Corporation (NYSE: XYZ), a leading global technology company, today announced that its Board of Directors has approved a stock buyback program. The program authorizes the repurchase of up to $500 million of the company's outstanding common stock over the next 12 months. John Doe, CEO of XYZ Corporation, stated, "This buyback program underscores our confidence in XYZ's long-term growth prospects and our commitment to delivering value to our shareholders. Our strong balance sheet and cash flow enable us to return capital to shareholders while continuing to invest in strategic initiatives."	1
Federal Reserve Announces Interest Rate Hike Washington, D.C. - June 14, 2024 - The Federal Reserve today announced an increase in the federal funds rate by 0.25 percentage points, raising the target range to 5.25-5.50%. This decision reflects the ongoing strength of the U.S. economy and aims to curb inflationary pressures. Jerome Powell, Chair of the Federal Reserve, stated, "Today's rate hike is a proactive measure to ensure that inflation remains in check while supporting sustained economic growth. We will continue to monitor economic indicators and adjust our policies as needed to maintain stability and promote maximum employment."	1
TechCorp Faces Backlash Over New Product Release San Francisco, CA - June 14, 2024 - TechCorp's latest smartphone, the X1000, has received significant criticism from consumers. Despite its high price tag of $1,199, the X1000's performance has fallen short of expectations, with many users reporting issues such as slow processing speeds and subpar battery life. Tech enthusiasts and early adopters have expressed their disappointment on social media, highlighting that the X1000 does not justify its premium price. Jane Doe, a tech analyst, commented, "TechCorp needs to address these performance issues quickly to regain consumer trust."	0
MegaCorp Engulfed in Embezzlement Scandal New York, NY - June 14, 2024 - MegaCorp, one of the world's largest conglomerates, is facing a major controversy following allegations of embezzlement. Reports indicate that several high-ranking executives are under investigation for misappropriating company funds amounting to millions of dollars. The scandal has sparked outrage among shareholders and the public, with many calling for immediate resignations and a thorough audit. In response, MegaCorp's CEO, John Smith, stated, "We are taking these allegations very seriously and have launched a comprehensive internal investigation. We are	0

committed to transparency and integrity in all our operations." For further information, please contact: MegaCorp Media Relations (555) 987-6543	