Sentiment analysis and stock price prediction integrated into Apple news/reviews

Data Sources

There are 3791 Weibo data (news/comments) related to Apple. Time span: 2012/11/19-2022/7/23.

Draw Word Cloud

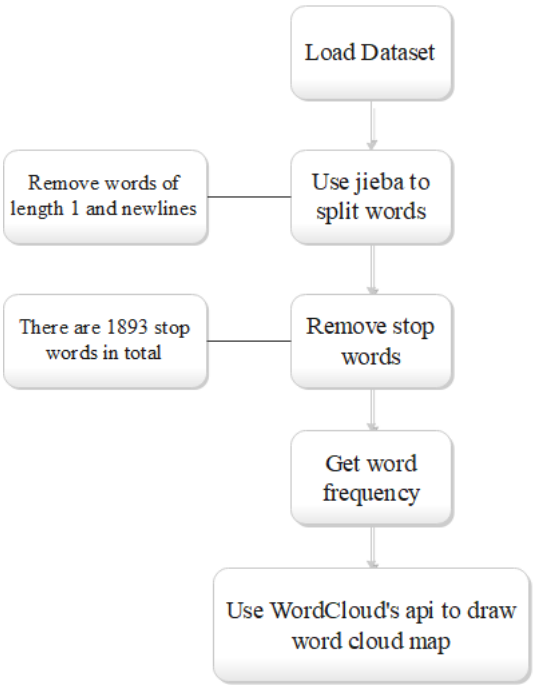

Figure 1: Flow Chart of Text Preprocessing.

time	contents
2012-12-31	【巴伦周刊：苹果股价仍被低估】巴伦周刊文章称，现在持有...
2012-12-31	【苹果股票市值3个月蒸发1750亿美元】苹果股票市值在...
2012-12-31	【苹果股票市值3个月蒸发1750亿美元】苹果股票市值在...
2012-12-31	【苹果股票再近期创新低：较峰值下跌27%】苹果股票今天...
2012-12-31	【苹果CEO库克今年薪酬缩水98.9%】据苹果周四提交...
2012-12-31	【财政悬崖前 分析师下调苹果目标股价】据外媒报道，最新...
2012-12-31	【苹果股价创新低 谷歌趁机"落井下石"】苹果股票再受重...
2012-12-31	【苹果股票市值3个月蒸发1750亿美元】北京时间12月...
2012-12-31	分享网易新闻:「美上调个税 苹果股票雪上加霜」 O网页...
2012-12-31	安趣新闻#【分析师认为苹果股票目标价的均值在740美元】 当某支股票极远地漂离目标价时，分析...

Figure 2: 10 Comments Data (example).

Figure 3: Apple style Word Cloud illustration.

Figure 1 presents the preprocessing of text data. According to statistics, there are 907

days of relevant news or comments, among which the words that appear more often are "股票", "投资", "下跌", "上涨" and so on. The darker the color in the word graph, the larger the size of the phrase, the more frequent it appears. Word cloud graphs can help us understand the content and topics of text more clearly.

Stock Market News Sentiment Analysis

SnowNLP is a class library written in python, dedicated to processing Chinese text. It can perform various tasks: part-of-speech tagging, sentiment judgment, keyword extraction, and generalizing text information. We mainly use this library to obtain sentiment scores of texts. The score is between 0 and 1, 0 means negative, one means positive, and 0.5 can be used to distinguish between positive and negative.

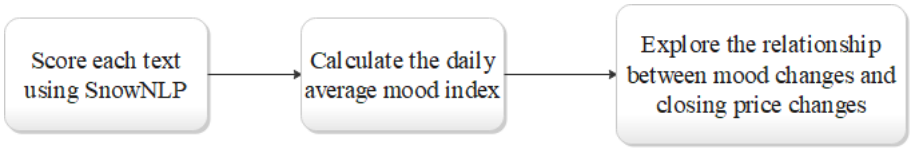

Figure 4: Flow chart of text preprocessing.

Date	Comments	Score	T/F
2012/12/31	【苹果股票市值 3 个月蒸发 1750 亿美元】苹果股票市值在过去 3 个月中蒸发了 1750 亿美元，创下了科技股之最，超过了惠普和 RIM 市值蒸发数额总和。就整个历史交易来说，惠普、RIM 和苹果的股票市值均遭遇了大幅缩水。数年前，惠普的股价曾达到了 50 美元以上的高点，而现在仅为 14 美元。	0.0795	T
2012/12/17	【iPhone 5 在中国首个周末销量创记录】苹果公司 12 月 17 日表示，自 12 月 14 日在中国内地发售 iPhone 5 的首周，销量突破了 200 万部。苹果公司首席执行官库克表示，中国客户对 iPhone 5 的需求令人惊讶，这是在中国销量最好的周末。花旗上周末下调苹果股票评级至“中性”，目标价从 675 美元下调至 575 美元。	0.9999	T
2012/11/28	【蒂姆·库克：最“波动”的接班人】之前被普遍看好的苹果接班人蒂姆· 库克最近的心情应该同公司股价一样在不断“波动”。乔布斯让苹果成为神话，用了 30 余年的时间，而在乔布斯去世一年多的今天，苹果股票从最高峰的每股 705.07 美元。	0.9998	T/F?
2017/8/15	万人迷的苹果和 Facebook 股票失去对冲基金偏爱 北京时间 15 日彭博称，对冲基金对 FAANG（Facebook、苹果、亚马逊、Netflix Inc．和谷歌母公司 Alphabet）的喜爱已经持续太久，无怪乎其中两家公司 -- 苹果和 Facebook 的吸引力正在下降。Ken Griffin 的 Citadel 卖出大部分苹果股票，减持数量达 340 万	0.9947	F

Table 1: Stock Comments Score (Example).

It can be seen that SnowNLP can correctly judge the sentiment index of some texts, but for the third and fourth rows in Table 1, the judgment is wrong due to the lack of text information and other reasons. However, manual annotation cannot be performed due to a large amount of data, so we use SnowNLP to turn unsupervised text into labeled data.

The relationship between sentiment index and the range of ups and downs(change).

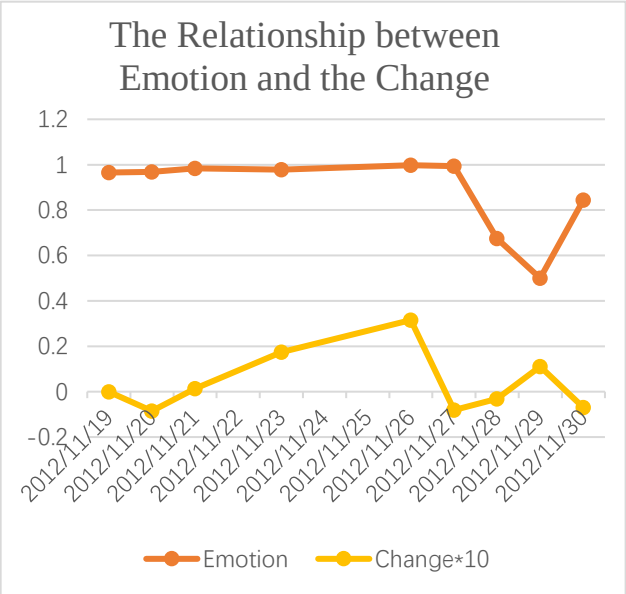

Figure 5: Sentiment Score and corresponding Closing Price Change (example 1).

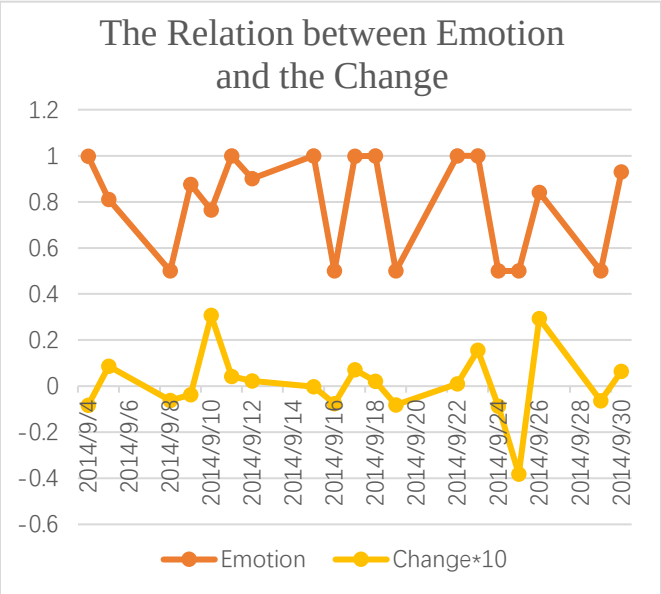

Figure 6: Sentiment Score and corresponding Closing Price Change (example 2).

Due to the extensive missing data, small total amount, and poor coherence, we extracted some coherent parts for comparison. We found that the rise and fall of Apple stock are related to the fluctuation of the sentiment index (Figures 5 and 6). However, there is a certain lag, and we cannot determine the optimal number of days for this lag because it may fluctuate.

We focused on the data from 2022-01-24 to 2022-03-30 (the longest complete, coherent data). We took ten as the window size and did correlation analysis on the **sentiment score (Emotion)** and the **closing price change (Change)**. Among them, we made Change lag behind Emotion by zero days, one day, two days, three days, and four days, and explore their corresponding correlation coefficients, respectively.

$$r(E,C) = \frac{Cov(E,C)}{\sqrt{Var[E]Var[C]}} \tag{1}$$

where $Cov(E,C)$ is the covariance of Emotion and Change, $Var[E]$ is the variance of Emotion, and $Var[C]$ is the variance of Change.

Date	Correlation Coefficient				
	4 days lag	**3 days lag**	2 days lag	1 day lag	0 day lag
2022/1/24	0.337	**0.424**	0.029	-0.145	0.132
...	...	**...**	...	...	...
2022/3/10	-0.491	**0.204**	-0.256	-0.099	-0.283
2022/3/11		**0.032**	-0.240	-0.049	-0.263
2022/3/14			-0.432	0.199	-0.215
2022/3/16				0.161	-0.358
2022/3/17					0.059
Mean(r)	-0.142	**0.315**	0.146	-0.161	-0.319

Table 2: Correlation coefficient versus lag days (window size 10).

Table 2 presents the correlation coefficient of Emotion and Change and the relationship between different lag days. It can be seen that when Change lags Emotion by three days, the mean value of the correlation coefficient is the largest, which is 0.315. Although this value is still insignificant, it can indicate that the sentiment index impacts the rise and fall of stocks and should be positively correlated.

Apple stock daily closing price forecast

Autoregressive integrated moving average (ARIMAX) models extend ARIMA models through the inclusion of exogenous variables X. We write an $ARIMAX(p, d, q)$ model for some time series data y_t and exogenous data X_t, where p is the number of autoregressive lags, d is the degree of differencing and q is the number of moving average lags as:

$$\Delta_{y_t}^{D} = \sum_{i=1}^{p} \emptyset_i \Delta_{y_{t-i}}^{D} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{m=1}^{M} \beta_m Xm, t + \epsilon_t \tag{2}$$

$$\epsilon_t \sim N(0, \sigma^2)$$

Visualize the daily closing price of Apple stock

We added the 5-day moving average, 10-day moving average and 20-day moving average of Apple's closing price. The simple moving average (SMA) over the last k days is calculated by Equation (3) as follows:

$$SMA_k = \frac{p_{n-k+1} + p_{n-k+2} + \cdots + p_n}{k} = \frac{1}{k} \sum_{i=n-k+1}^{n} p_i \tag{3}$$

where p_i is the value of the gold/bitcoin on i-th day.

Figure 7: Apple's daily closing price and EM_5, EM_10, EM_20 (SMA).

Figure 7 presents Apple's daily closing curve and corresponding moving average. When the short-term SMA(EM_5) exceeds the long-term SMA(EM_20), the asset has an upward trend, implying that it is suitable for buying. When the short-term SMA moves down and intersects with the long-term SMA, the asset has a downward trend, making it suitable for selling.

Construction of ARIMAX Model

Symbol	Definition
Close	Daily closing price of Apple stock.
Adj_Close	Adjusted daily closing price of Apple stock.
Volume	Apple Stock Daily Volume.
Emotion	Daily Apple stock sentiment index.
Change	The difference between the closing prices of the stock for the previous 2 days.
Close_MA_5	Average of closing prices for the previous 5 days.
Close_MA_10	Average of closing prices for the previous 10 days.
Close_MA_20	Average of closing prices for the previous 20 days.

Table 3: Input variables and interpretation of the ARIMAX model.

First, we used the ADF test (Augmented Dickey-Fuller test) for the stationarity of Apple's closing price. The ADF test determines whether the sequence has a unit root: if the sequence is stationary, there is no unit root. Otherwise, there is a unit root.

Test Statistic	0.494763
p-value	0.984713

Lags Used	27
Number of Observations Used	2387
Critical Value (1%)	-3.433092
Critical Value (5%)	-2.862752
Critical Value (10%)	-2.567415

Table 4: Apple Closing Curve Results.

The p-value in Table 4 is greater than 0.05, indicating that the series is volatile. Next, we performed first-order differences and second-order differences on the original data. After finding the first-order difference, we performed the ADF test to meet the requirements (p-value is 4.871830e-16).

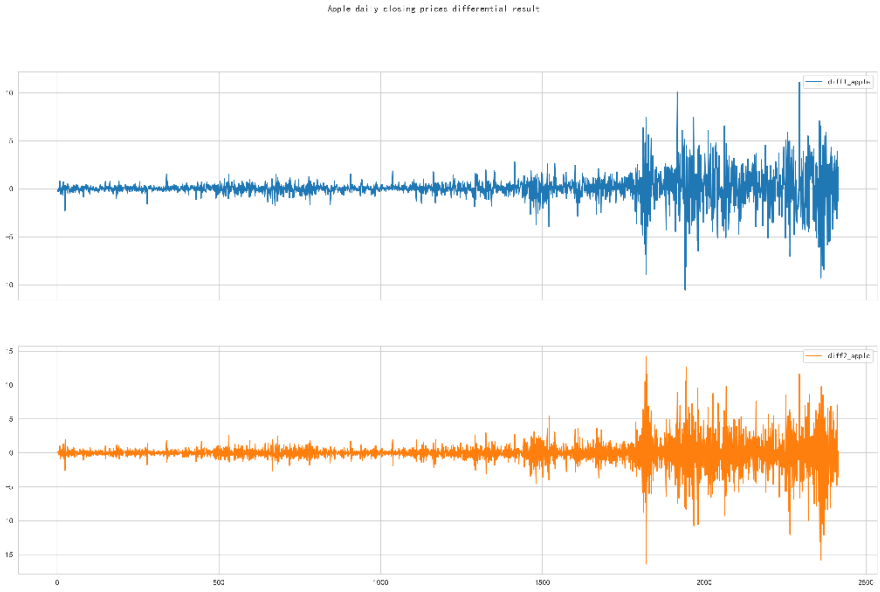

Figure 8: Apple daily closing prices differential result.

We determined the model's parameters by ACF (Autocorrelation function) and PACF (Partial Autocorrelation function). ACF describes the autocorrelation between one observation and another, including direct and indirect correlation information. Instead of finding the correlation of a lag like the ACF with the current, it finds the correlation of the residual (which remains after removing the effects already explained by the previous lags) with the value of the next lag.

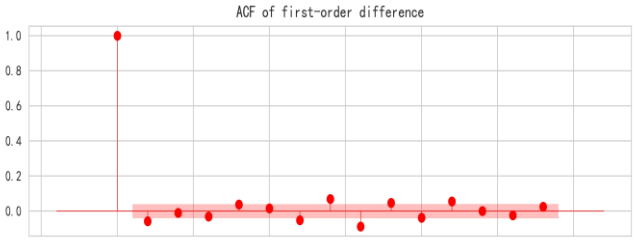

Figure 9: ACF.

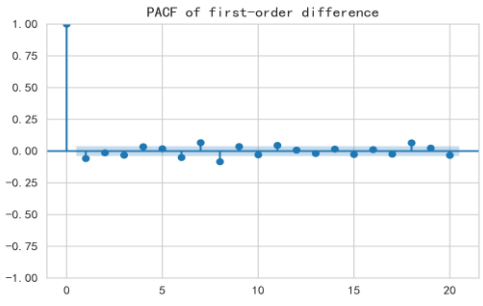

Figure 10: PACF.

From Figures 9 and 10, it was preliminarily judged that d=1 (first-order difference) in ARIMAX(p,d,q), p=q=2 (or 1) in $ARIMAX(p,d,q)$. We used the first 75% of the dataset as the training set and the last 25% of the dataset as the validation set to start training.

	Patsy Formula	AR	MA	integ	AIC	BIC
1	**Close~0+Emotion+Change+Adj_Close+Close_MA_5 +Close_MA_10+Close_MA_20+Adj_Close**	**1**	**1**	**1**	**-109.80**	**-60.30**
2	Close~0+Chang	1	1	1	-9.86	12.14
3	Close~Change	1	1	1	-7.91	19.59
4	Close~Change	1	2	1	-6.12	26.88
5	Close~Change	2	1	1	-5.75	27.24
6	Close~Change	2	2	1	-4.36	34.13
7	Close~0+Change+Adj_Close+Close_MA_5+ Close_MA_10+Close_MA_20+Adj_Close	1	1	1	2664.21	2708.21
8	Close~Close	1	1	1	3251.01	3278.51
9	Close~Adj_Close	1	1	1	3254.66	3282.15
10	Close~Emotion	1	1	1	3260.46	3287.95
11	Close~Close_MA_10	1	1	1	3260.57	3288.07
12	Close~Close_MA_5	1	1	1	3260.62	3288.12
13	Close~Close_MA_5	1	1	1	3261.18	3288.68
14	Close~Change	1	1	2	3465.90	3493.40
15	Close~Volume	1	1	1	3972.13	3999.62

Table 5: Evaluation results of ARIMAX under different subordinates.

AIC (Akaike information criterion) is a standard to measure the goodness of statistical model fitting. The smaller the AIC, the better the model, and the model with the smallest AIC is usually selected. It can be seen that the parameter configuration in the first line is optimal.

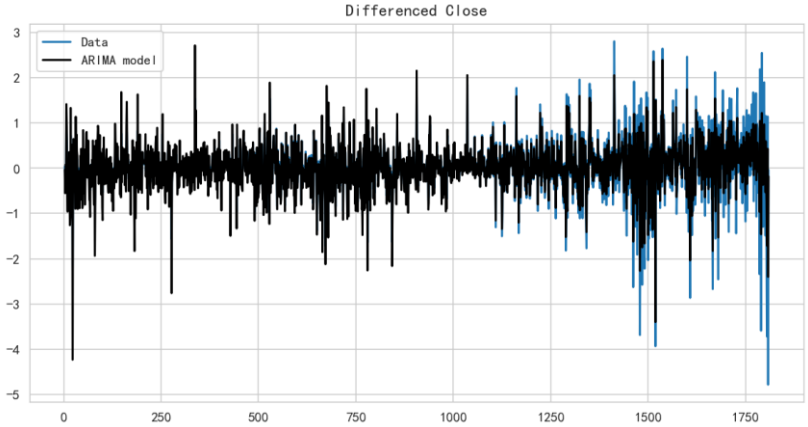

Figure 11: Fitting of ARIMAX(1,1,1) on the training set.

We select 3 indicators as evaluation criteria for model performance—mean square

error (MSE), root mean square error ($RMSE$) and efficient of determination (R^2).

The mean square error (MSE) is calculated by Equation (4) as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

where y_i is the true value, and $\hat{y}_i$ is the predicted value of y_i.

The root mean square error ($RMSE$) is calculated by Equation (5) as follows:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{5}$$

where MSE is the mean square error.

The coefficient of determination (R^2) is calculated by Equation (6) as follows:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

where SS_{res} is the sum of squares of residuals, SS_{tot} is the total sum of squares, y_i is the true value, $\bar{y}$ is the mean of true values, and $\hat{y}_i$ is the predicted value of y_i.

MSE	3.607
RMSE	1.899
R^2	0.996

Table 6: Performance of ARIMAX(1,1,1) on the test dataset.

It can be seen that this is a good result, which confirms that adding sentiment analysis and financial indicators such as EM can show a better prediction effect to a certain extent. However, a severe hysteresis problem still needs to be solved. The lag in stock price forecasts can lead to poor investment outcomes.