

Natural Language Processing

Yuen Tak Cheung 57160006

Keung Yat Long 57146792

Problem Formulation

- **Comment** of social media may affect **stock price** of a company
- Analysing **textual content** from social media may brought us **insight**
- **Prediction** may be done by the **relationship** between **comment** and **stock price**
- Case study: **Tesla** as the subject of analysis

Datasets

Financial dataset

Date	Open	High	Low	Close	Adj Close	Volume	Stock Name
30/9/2021	260.333344	263.043335	258.333344	258.493347	258.493347	53868000	TSLA
1/10/2021	259.466675	260.26001	254.529999	258.406677	258.406677	51094200	TSLA
4/10/2021	265.5	268.98999	258.706665	260.51001	260.51001	91449900	TSLA
5/10/2021	261.600006	265.769989	258.066681	260.196655	260.196655	55297800	TSLA
6/10/2021	258.733337	262.220001	257.73999	260.916656	260.916656	43898400	TSLA
7/10/2021	261.820007	268.333344	261.126678	264.536682	264.536682	57587400	TSLA
8/10/2021	265.40332	265.459991	260.303345	261.829987	261.829987	50215800	TSLA

30/9/2021	285.709991	287.829987	281.619995	281.920013	278.792847	32343600	MSFT
1/10/2021	282.119995	289.980011	281.290009	289.100006	285.893219	30086300	MSFT
4/10/2021	287.399994	287.75	280.25	283.109985	279.969635	31350700	MSFT
5/10/2021	284.049988	290.399994	284.049988	288.76001	285.556946	24993000	MSFT
6/10/2021	285.779999	293.630005	285.51001	293.109985	289.858673	28002600	MSFT
7/10/2021	295.179993	296.640015	293.920013	294.850006	291.579437	20430500	MSFT
8/10/2021	296.220001	296.640015	293.76001	294.850006	291.579437	17685700	MSFT

- Provide historical stock market data

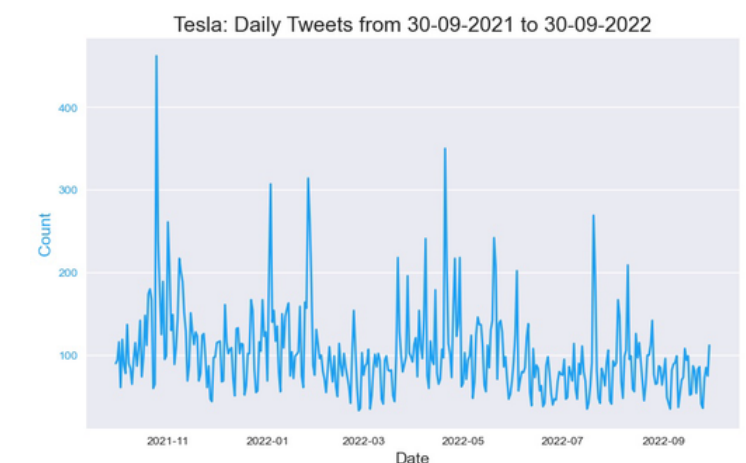


Tweets dataset

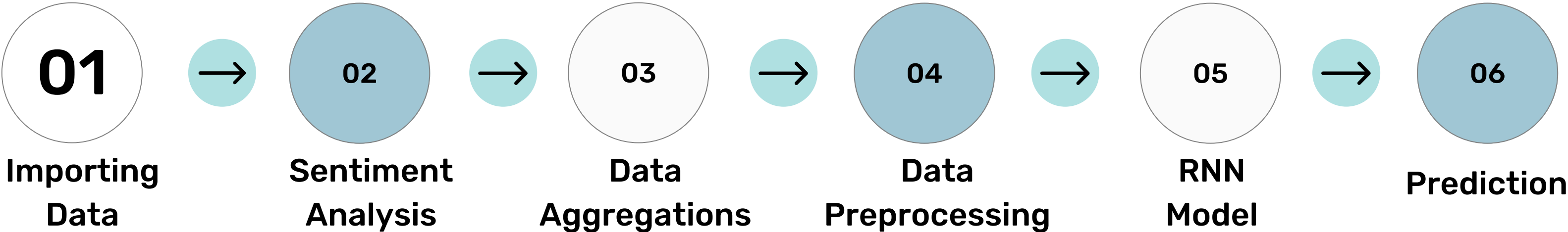
Date	Tweet	Stock Name	Company Name
2022-09-29 23:41:16+0	Mainstream media has done an amazing job at brainwashing people. Today at	TSLA	Tesla, Inc.
2022-09-29 23:24:43+0	3/ Tesla delivery estimates are at around 364k from the analysts. \$tsla	TSLA	Tesla, Inc.
2022-09-29 23:18:08+0	3/ Even if I include 63.0M unvested RSUs as of 6/30, additional equity needed	TSLA	Tesla, Inc.
2022-09-29 22:40:07+0	@RealDanODowd @WholeMarsBlog @Tesla Hahaha why are you still trying t	TSLA	Tesla, Inc.
2022-09-29 22:27:05+0	@RealDanODowd @Tesla Stop trying to kill kids, you sad deranged old man	TSLA	Tesla, Inc.
2022-09-29 22:25:53+0	@RealDanODowd @Tesla This is you https://t.co/3Mt1XawSEb	TSLA	Tesla, Inc.
2022-09-29 22:24:22+0	For years @WholeMarsBlog viciously silenced @Tesla critics. Failing to silen	TSLA	Tesla, Inc.
2022-09-29 22:23:54+0	\$NIO just because I'm down money doesn't mean this is a bad investment. The	TSLA	Tesla, Inc.

30/9/2021	285.709991	287.829987	281.619995	281.920013	278.792847	32343600	MSFT
1/10/2021	282.119995	289.980011	281.290009	289.100006	285.893219	30086300	MSFT
4/10/2021	287.399994	287.75	280.25	283.109985	279.969635	31350700	MSFT
5/10/2021	284.049988	290.399994	284.049988	288.76001	285.556946	24993000	MSFT
6/10/2021	285.779999	293.630005	285.51001	293.109985	289.858673	28002600	MSFT
7/10/2021	295.179993	296.640015	293.920013	294.850006	291.579437	20430500	MSFT
8/10/2021	296.220001	296.640015	293.76001	294.850006	291.579437	17685700	MSFT

- Tweets related to stock market discussion




Method of Approach



Date	Tweet	StockName
2022-09-29 23:41:16+0	Mainstream media has done	TSLA
2022-09-29 23:24:43+0	Tesla delivery estimates are	TSLA
2022-09-29 23:18:08+0	3/ Even if I include 63.0M unv	TSLA
2022-09-29 22:40:07+0	@RealDanODowd @Wholeh	TSLA
2022-09-29 22:27:05+0	@RealDanODowd @Tesla St	TSLA
2022-09-29 22:25:53+0	@RealDanODowd @Tesla Th	TSLA
2022-09-29 22:24:22+0	For years @WholeMarsBlog	TSLA
2022-09-29 22:23:54+0	\$NIO just because I'm down	TSLA


Date	Close
30/9/2021	258.493347
1/10/2021	258.406677
4/10/2021	260.51001
5/10/2021	260.196655
6/10/2021	260.916656
7/10/2021	264.536682
8/10/2021	261.829987
11/10/2021	263.980011
12/10/2021	268.573334
13/10/2021	270.359985
14/10/2021	272.773346
15/10/2021	281.01001
18/10/2021	290.036682
19/10/2021	288.089996
20/10/2021	288.600006
21/10/2021	298

sentiment_score	Negative	Neutral	Positive
0.0772	0.127	0.758	0.115
0.0	0.0	1.0	0.0
0.296	0.0	0.951	0.049
-0.7568	0.273	0.59	0.137
-0.875	0.526	0.474	0.0




POSITIVE

"Great service for an affordable price. We will definitely be booking again."



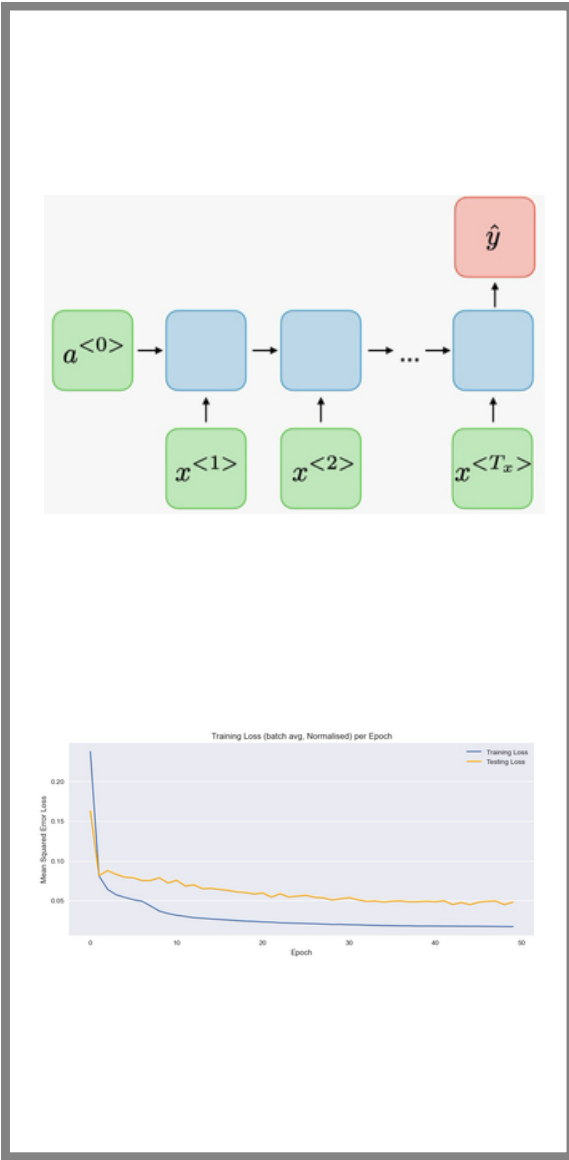
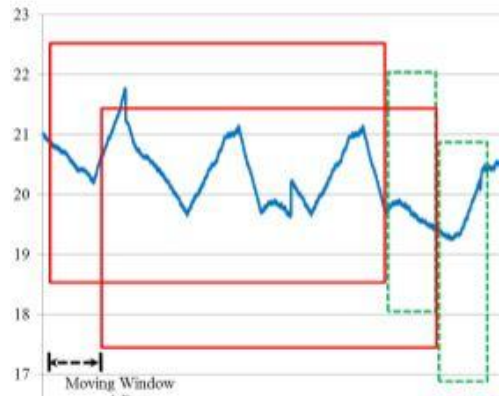
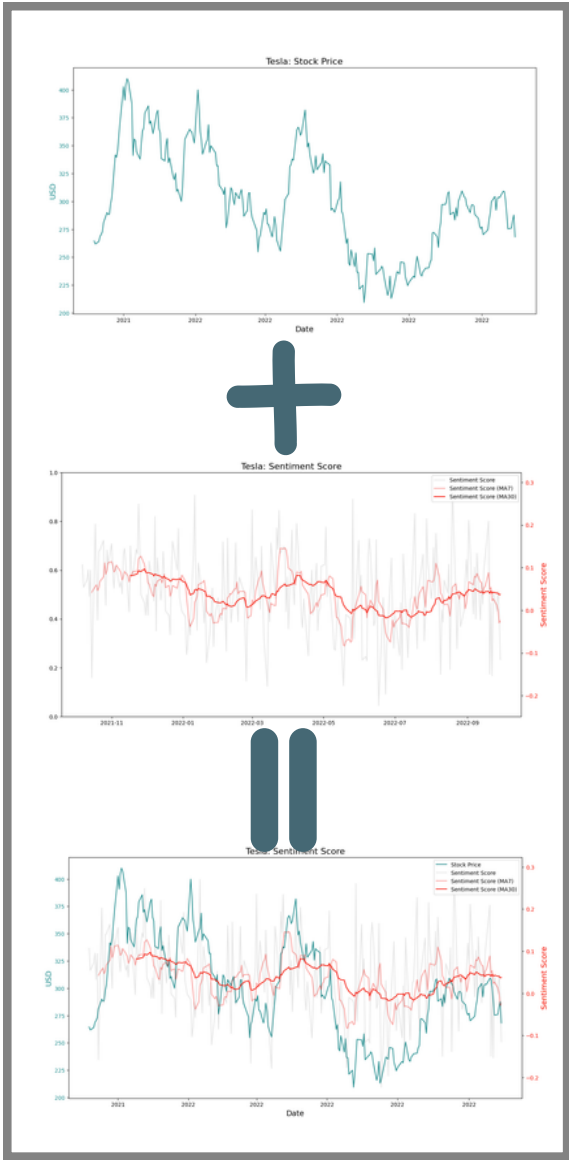
NEUTRAL

"Just booked two nights at this hotel."



NEGATIVE

"Horrible services. The room was dirty and unpleasant. Not worth the money."



NLP Model for Sentiment Analysis

VADER: lexicon-based NLP model

- Strong words have very positive or negative sentiment rating
- Polarity score: sum of the ratings, normalised to range (-1,1)

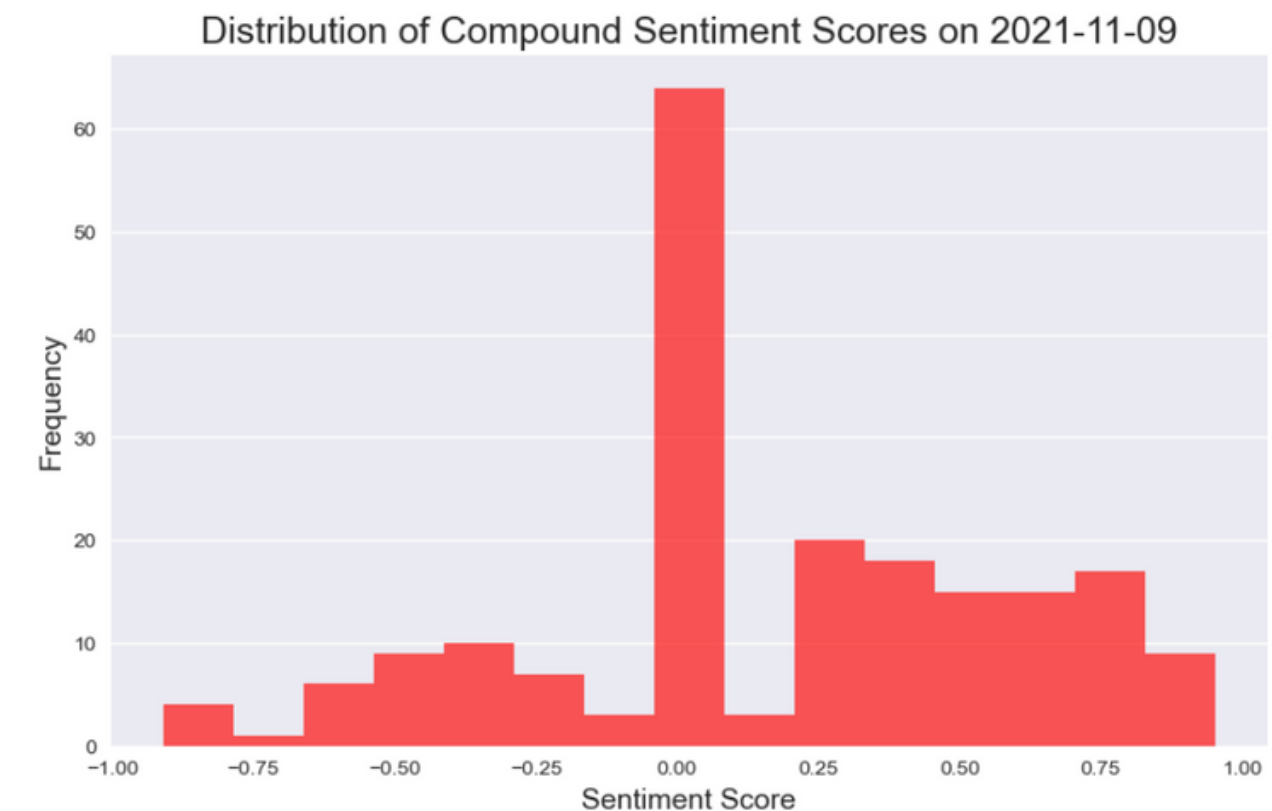
Word	Sentiment rating
tragedy	-3.4
rejoiced	2.0
insane	-1.7
disaster	-3.1
great	3.1

$$x = \frac{x}{\sqrt{x^2 + \alpha}}$$

where x = sum of valence scores of constituent words, and α = Normalization constant (default value is 15)

Why choose rule-based model over embedding-based model for Tweets?

- Polarity score more informative for inputs than positive/negative
- Recognise evolving language: emojis (ð_ð) and slang (Sus)
- Very fast with high accuracy



(S.D. = 0.43-0.46 per day)

Data Aggregations

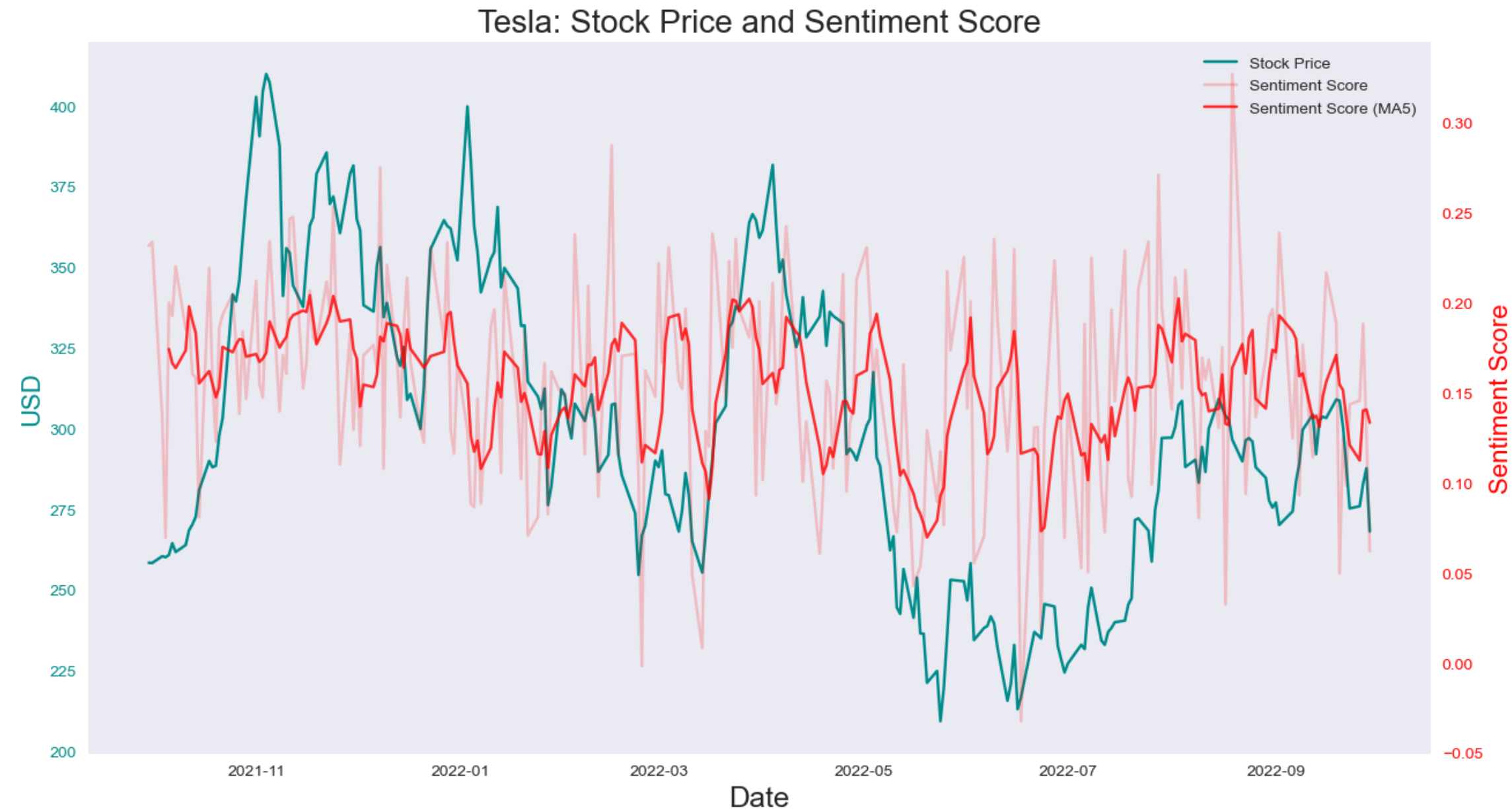
Further text cleaning

Some special cases which are not recognised by VADER need manual removal:

- URLs: “http://great.china.mobile.com”
- Mentions (usernames): @terrible_peter

Relating sentiments on stock prices

- Group sentiment scores by dates and then take averages
- Ignore weekends' **tweets** when the market is closed
- At least one month of dependency (**20 days**) is needed



Data Preprocessing

Prepare inputs and outputs

1. Split into training & testing sets
2. MinMaxScaler (0 to 1) for equal feature contribution
3. Sequential inputs & labels using window slicing

PyTorch Wrapper

1. Convert into Tensor
2. Batching using Dataloader for more efficient convergence

Entity	Shape	Type	Remark
dataset2	252x2	DataFrame	Close & sentiment_score
dataset_train, scaled_train	(202, 2)	Array	50 days for testing
dataset_test, scaled_test	(50, 2)	Array	
X_train, y_train	(162, 40, 2)	Array	X: ith day to (i+39)th day y: (i+1)th day to (i+40)th day
X_test, y_test	(30, 20*, 2)	Array	
batch_X, batch_y	(16, 40, 2)	Tensor	Batch size = 16
batch_X_test, batch_y_test	(16, 20, 2)	Tensor	

**Time sequence length for testing is not necessarily equal to training*

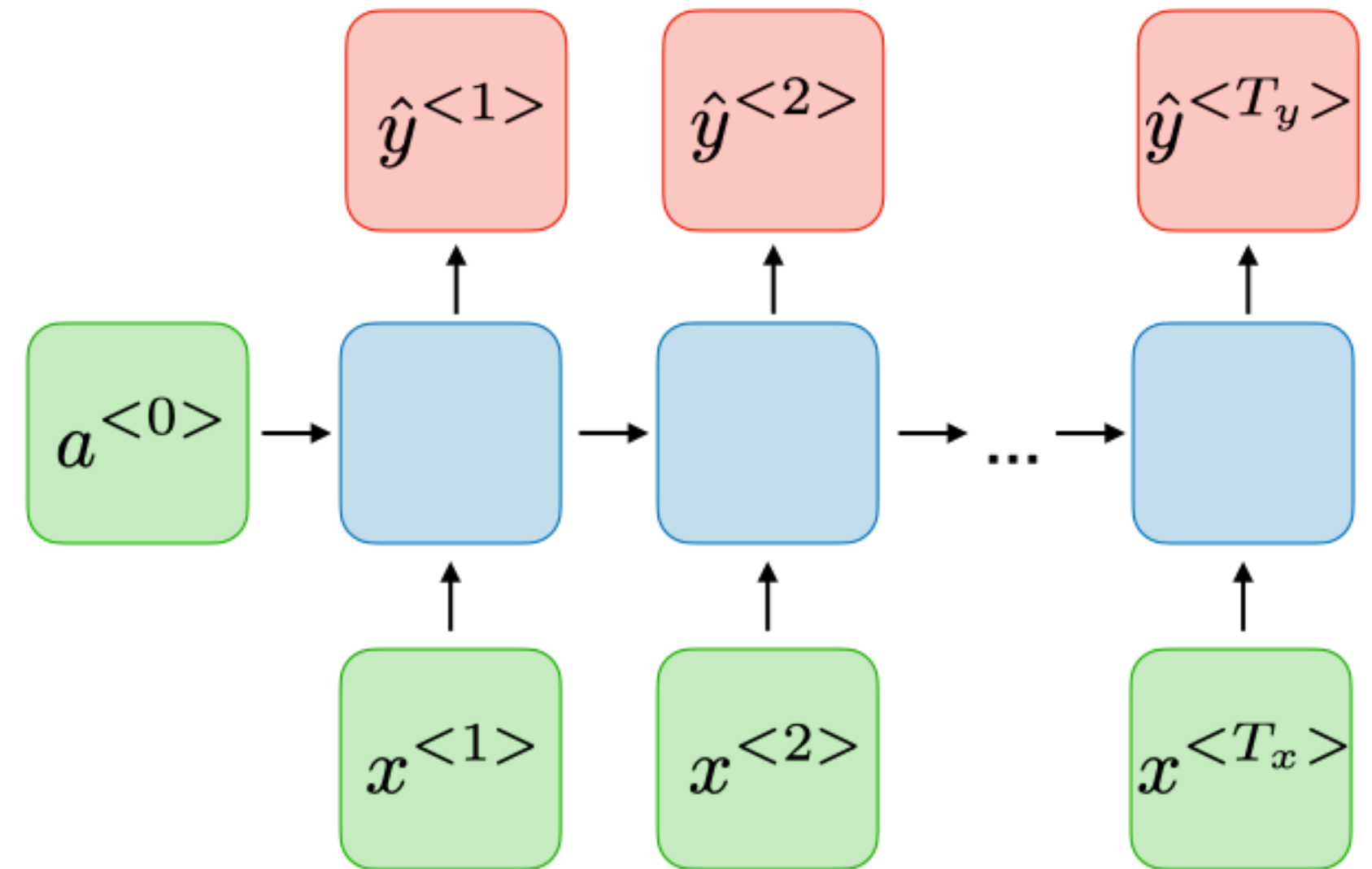
Recurrent Neural Network

Many-to-many architecture

- **Customisable** input and output lengths to predict future stock prices and sentiment scores
- **Shared weights** to generalise across different time steps and **recursive hidden states** to maintain memory of past inputs for context

Adjustments

- Long short-term memory (LSTM) network for **longer dependencies** and **prevents gradient vanishing**
- Adam Optimiser for **adaptive learning rate** (larger steps in simpler regions and smaller in more complex ones)
- **Randomly dropping** hidden states to avoid overfitting to specific patterns



Results

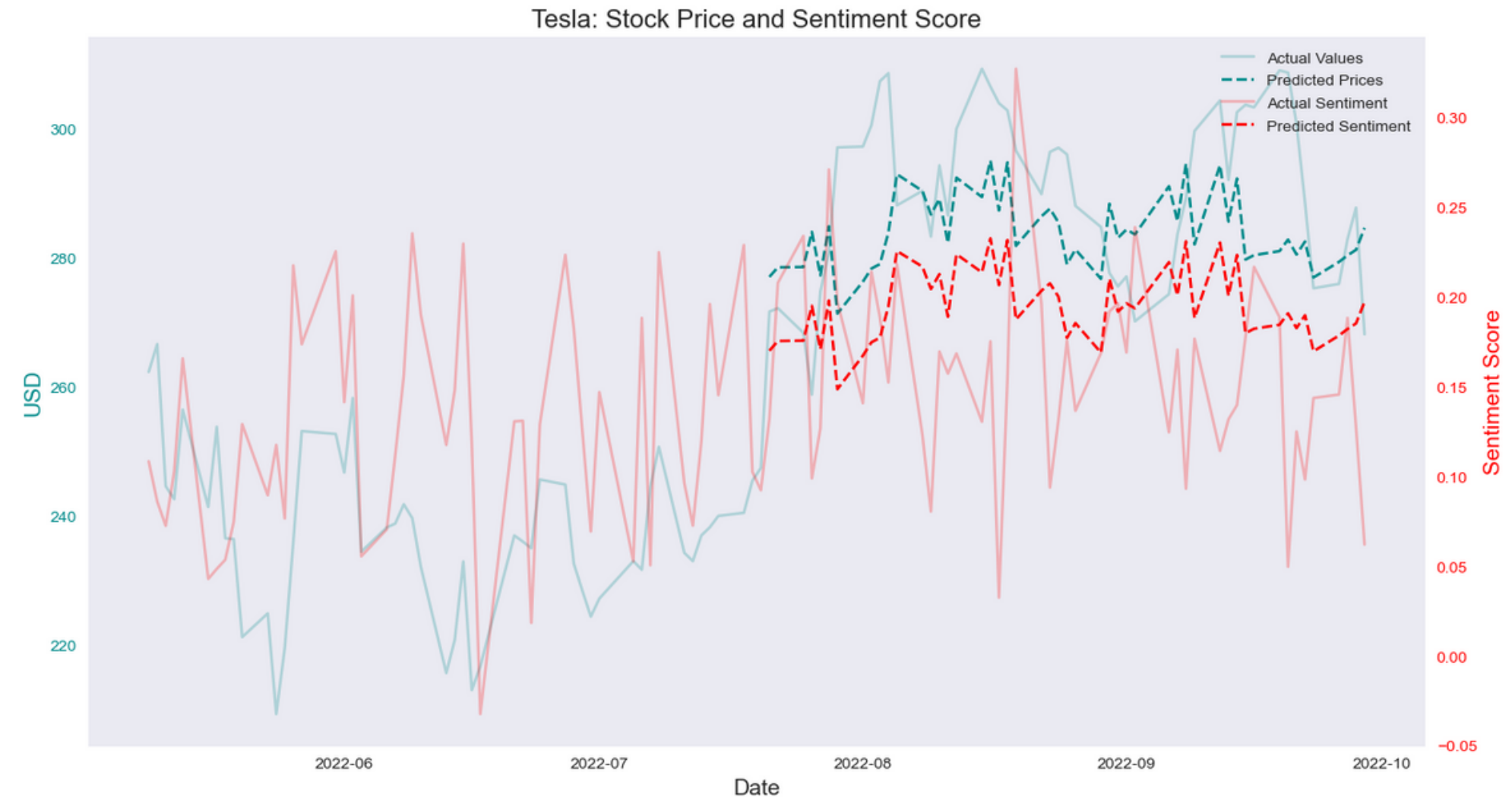
Conclusion: **Oversimplified** Model and **Insufficient** Features

Training



- Training temporarily stopped during evaluation
- Both losses **converged**
- High testing loss due to “residual batch” ($30 \div 16$)

Prediction



- Rolling Prediction based on past predictions
- Stock prediction follows actual trends with a **high bias**
- Sentiment prediction has a **very similar pattern**

Discussion

Challenging Issues

- Uncommon emojis and **ironic** sentences not recognised:

Cleaned_Tweet	sentiment_score
Can't wAlt for AI Day 2. 🤖 \$TSLA	0.0
"I emailed IR to say Elon needs to shut up." \$...	0.0
People selling everything to buy the \$TSLA dip.	0.0

- Outer **uncertainties**: Elon Musk & COVID shutdown:



Tesla's new Model 3 sedan displayed at the China International Fair for Trade in Services (CIFTIS) in Beijing, China, Sept. 2, 2023. (Florence Lo/REUTERS) (REUTERS / Reuters)

The Chinese Lunar Holiday, during which the country shuts down for nearly two weeks, fell in February of this year. Historically, this has led

Future Improvements

- Statistics indicators (Moving Average Convergence Divergence) & economic indicators (interest rates)
- Separate training for sentiments (with careful tuning)
- Multi-task learning with separate layers
- Extract more data directly from X API (**subscription**):

Developer Portal

TC

Dashboard

Projects

Sentiment&NLP

MONTHLY POST CAP USAGE

0 Posts pulled of 1,500

Resets on May 11 at 00:00 UTC

0%

```
try:
    tweets = api.search_tweets(
        attributes_container =
        columns = ["User", "Data"]
        tweets_df = pd.DataFrame(attributes_con
except BaseException as e:
    print("Status Failer On,", str(e))
```

✓ 0.5s

int, you may need a different access level. You