Textual Analysis for Stock Price Prediction

Group 11: Hongzhen Du, Zhengyuan Tan, Le Xu, Qiyao Xue

SWS Summer Workshop 2022
July 28th, 2022

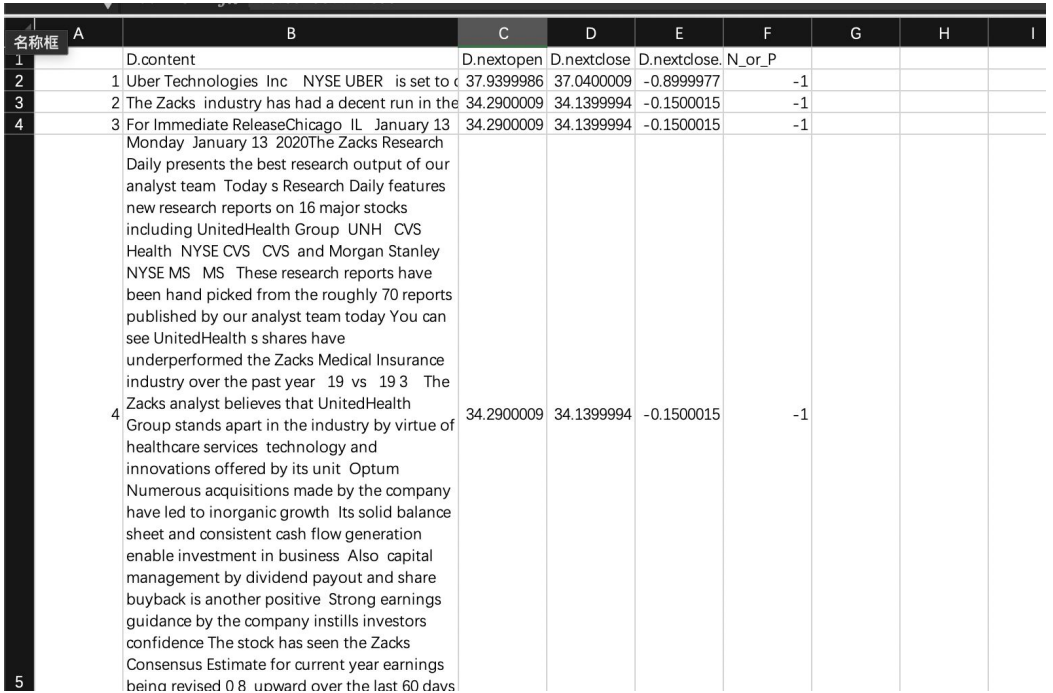
Data Preparation

R language & R studio

Stock-related news texts, Date for the news and Price (Open & Close)

Extract final data in an csv document

The First Glance to Our Dataset



	A	B	C	D	E	F	G	H	I
1		D.content	D.nextopen	D.nextclose	D.nextclose.	N_or_P			
2	1	Uber Technologies Inc NYSE UBER is set to	37.9399986	37.0400009	-0.8999977	-1			
3	2	The Zacks industry has had a decent run in the	34.2900009	34.1399994	-0.1500015	-1			
4	3	For Immediate ReleaseChicago IL January 13	34.2900009	34.1399994	-0.1500015	-1			
5	4	Monday January 13 2020The Zacks Research Daily presents the best research output of our analyst team Today s Research Daily features new research reports on 16 major stocks including UnitedHealth Group UNH CVS Health NYSE CVS CVS and Morgan Stanley NYSE MS MS These research reports have been hand picked from the roughly 70 reports published by our analyst team today You can see UnitedHealth s shares have underperformed the Zacks Medical Insurance industry over the past year 19 vs 19 3 The Zacks analyst believes that UnitedHealth Group stands apart in the industry by virtue of healthcare services technology and innovations offered by its unit Optum Numerous acquisitions made by the company have led to inorganic growth Its solid balance sheet and consistent cash flow generation enable investment in business Also capital management by dividend payout and share buyback is another positive Strong earnings guidance by the company instills investors confidence The stock has seen the Zacks Consensus Estimate for current year earnings being revised 0 8 upward over the last 60 days	34.2900009	34.1399994	-0.1500015	-1			

NLP Analysis Using a Lexicon

NRC Emotion Lexicon

Transform dataset into a simple one

Construct an English lexicon

Match the lexicon and calculate the score

Reduction of dataset

```
def depure_data(data):

    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    data = url_pattern.sub(r'', data)
    data = re.sub('\S*@\S*\s?', '', data)
    data = re.sub('\s+', ' ', data)
    data = re.sub("\'", "", data)

    return data

def sent_to_words(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence),deacc=True))

def detokenize(text):
    return TreebankWordDetokenizer().detokenize(text)
```

Refence:https://towardsdatascience.com/an-easy-tutorial-about-sentiment-analysis-with-deep-learning-and-keras-2bf52b9cba91

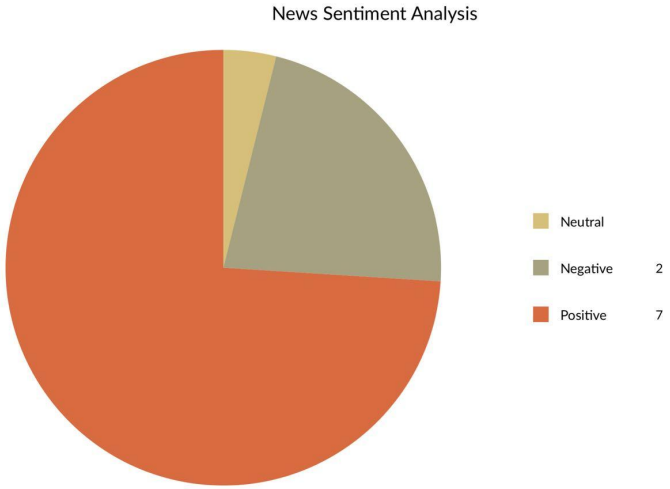
Construct an English lexicon

```
english_dic = lexion_df[['English (en)','Positive','Negative','Anger','Anticipation','Disgust','Fear','Joy','Sadness','Surprise','Trust']]
Positive = []
Negative = []
Anger = []
Anticipation = []
Disgust = []
Fear = []
Joy = []
Sadness = []
Surprise = []
Trust = []
for idx, row in english_dic.iterrows():
    if row ['Positive']==1:
        Positive.append(row['English (en)'])
    if row ['Negative']==1:
        Negative.append(row['English (en)'])
    if row ['Anger']==1:
        Anger.append(row['English (en)'])
    if row ['Anticipation']==1:
        Anticipation.append(row['English (en)'])
    if row ['Disgust']==1:
        Disgust.append(row['English (en)'])
    if row ['Fear']==1:
        Fear.append(row['English (en)'])
    if row ['Joy']==1:
        Joy.append(row['English (en)'])
    if row ['Sadness']==1:
        Sadness.append(row['English (en)'])
    if row ['Surprise']==1:
        Surprise.append(row['English (en)'])
    if row ['Trust']==1:
        Trust.append(row['English (en)'])
```

Result

Unnamed: 0	D. content	D. nextope	D. nextclo	D. nextclo	N_or_P	positive	negative	anger	anticipat	disgust	fear	joy	sadness	surprise	trust	length	sentiment	dependence
1	Uber Tech	37.939999	37.040001	-0.899998	-1	42	32	20	26	1	23	15	23	1	18	2926	1	0
2	The Zacks	34.290001	34.139999	-0.150002	-1	109	20	2	47	0	2	19	2	1	29	3544	1	0
3	For Immed	34.290001	34.139999	-0.150002	-1	181	89	11	40	3	24	24	19	9	81	9279	1	0
4	Monday January 13 2020The Zacks Research Daily presents the best research output of our analyst team Today s Research Daily features new research reports on 16 major stocks including UnitedHealth Group UNH	34.290001	34.139999	-0.150002	-1	739	73	25	397	2	24	15	19	4	369	6879	1	0
5	For Immed	34.290001	34.139999	-0.150002	-1	169	61	6	109	0	9	13	7	3	108	5031	1	0

Result



33.3%

Dependence

LSTM Model

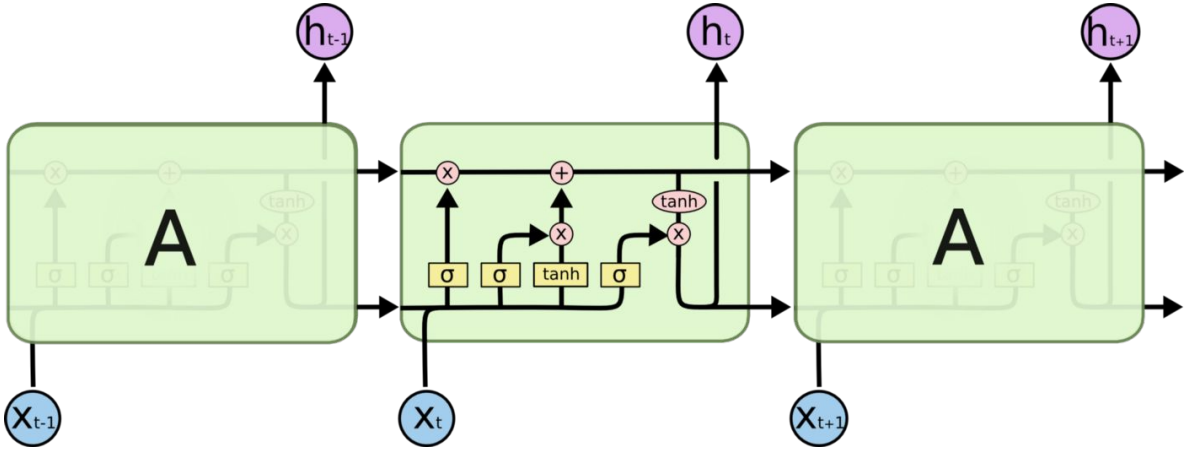
Python Tensorflow

Training set's accuracy is good, around 75%

Accuracy for out-sample dataset is at the average (below 50%).

Loss is decreasing / accuracy is increasing.

LSTM Model Main Structure



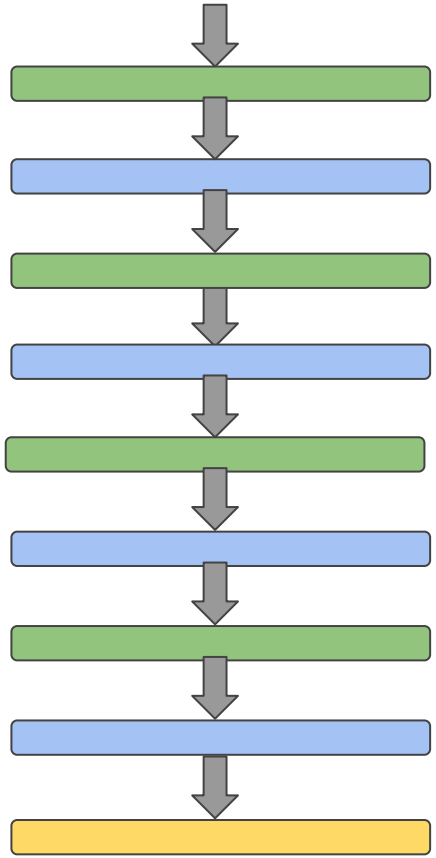
Design of the LSTM model for stock prediction

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 100, 50)	10400
dropout (Dropout)	(None, 100, 50)	0
lstm_1 (LSTM)	(None, 100, 60)	26640
dropout_1 (Dropout)	(None, 100, 60)	0
lstm_2 (LSTM)	(None, 100, 80)	45120
dropout_2 (Dropout)	(None, 100, 80)	0
lstm_3 (LSTM)	(None, 120)	96480
dropout_3 (Dropout)	(None, 120)	0
dense (Dense)	(None, 1)	121

Total params: 178,761
Trainable params: 178,761
Non-trainable params: 0

```
model = Sequential()
model.add(LSTM(units=50, activation='relu', return_sequences=True,
               input_shape=(x_train.shape[1], 1)))
model.add(Dropout(0.2))
model.add(LSTM(units=60, activation='relu', return_sequences=True))
model.add(Dropout(0.3))
model.add(LSTM(units=80, activation='relu', return_sequences=True))
model.add(Dropout(0.4))
model.add(LSTM(units=120, activation='relu'))
model.add(Dropout(0.5))

model.add(Dense(units=1))
```



LSTM:

ReLU+Dropout:

Fully connected:

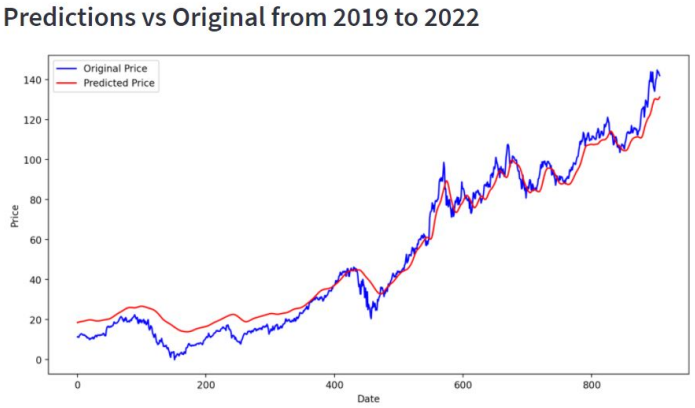
Optimizer:
Adaptive Moment Estimation

Loss function:
Mean Square Error

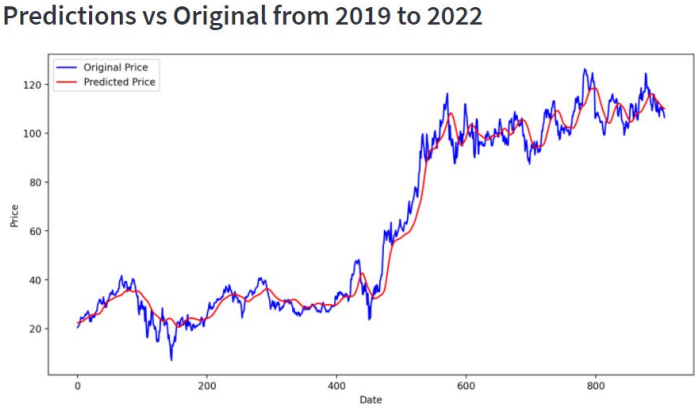
Eopch number:
50

Prediction result of the LSTM model

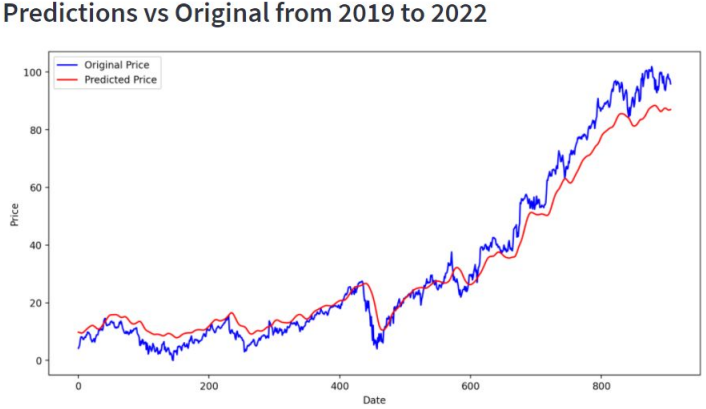
APPL



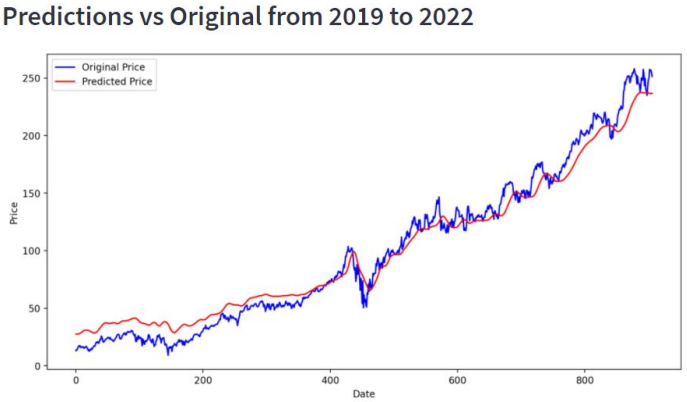
AMZN



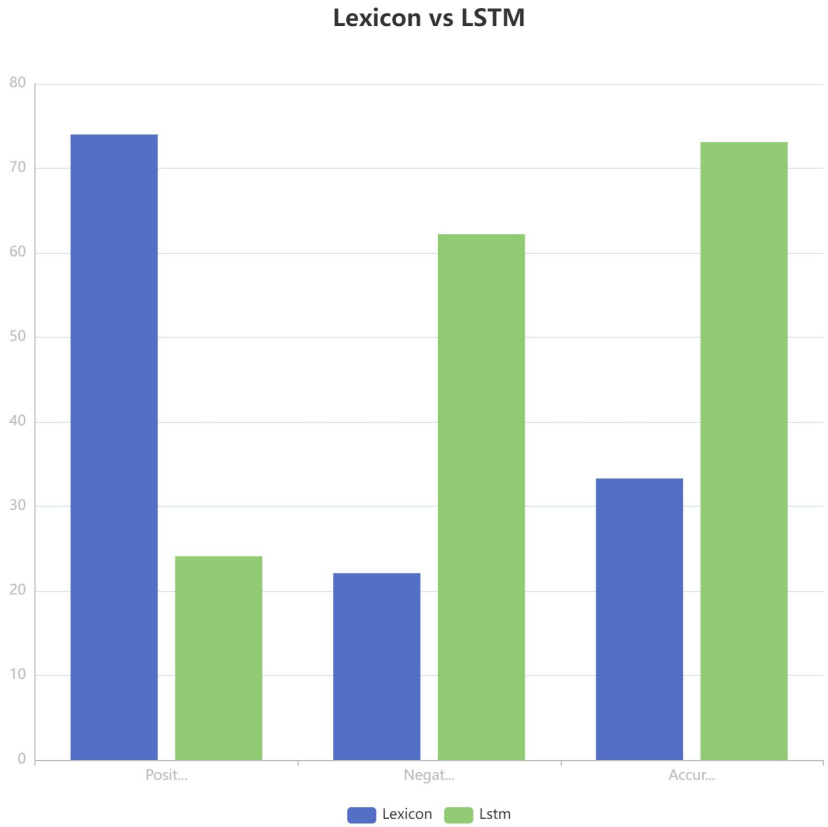
GOOG



MSFT

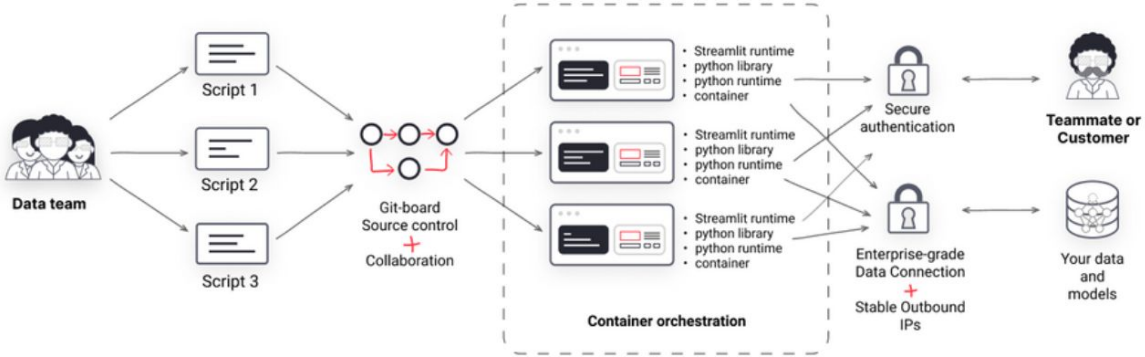


Compare with the lexicon analysis



Design of the whole project

Mainly developed on Stramlit based on python

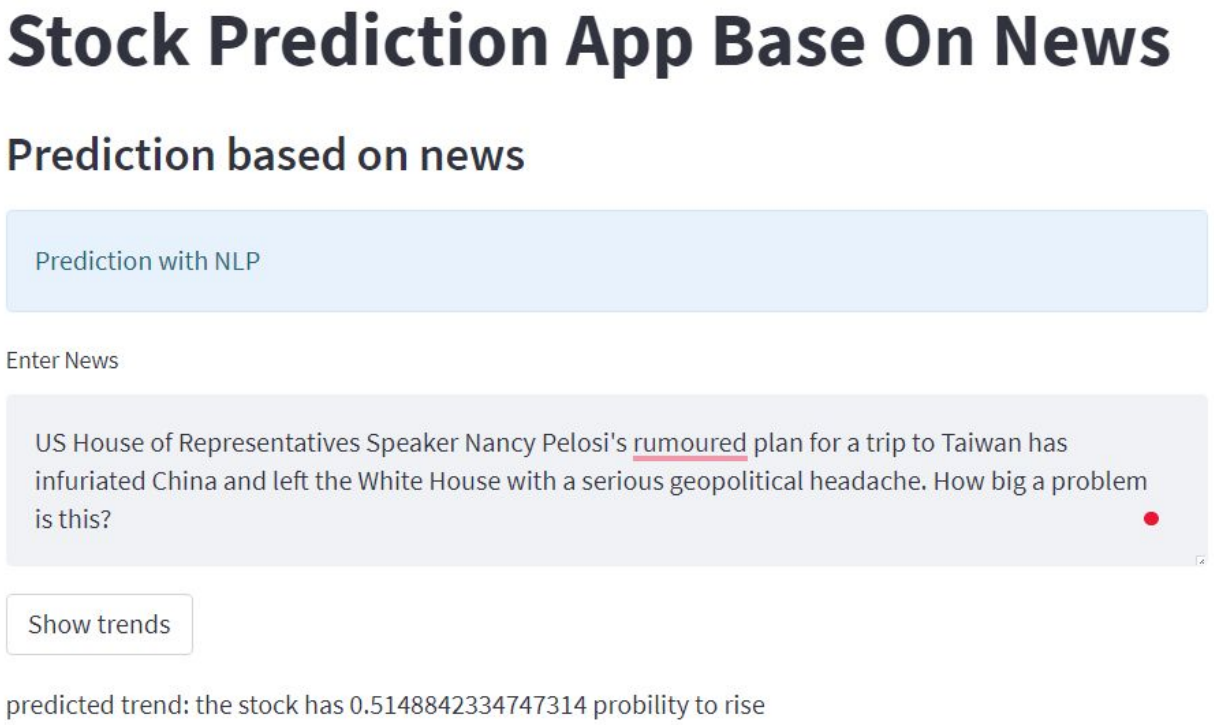


Streamlit

Combined with two parts Stock Precition and News NLP Analysis

Trend prediction based on news content

The app will show the influence of the news on the stock by returning the rising probility of the stock



Trend prediction based on previous price

The app will predict the future trend of stock based on the previous price and LSTM model

Prediction based previous price

Prediction with LSTM

Choose stock

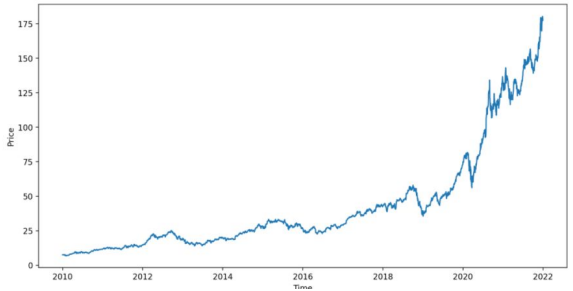
AAPL

Plot graphs

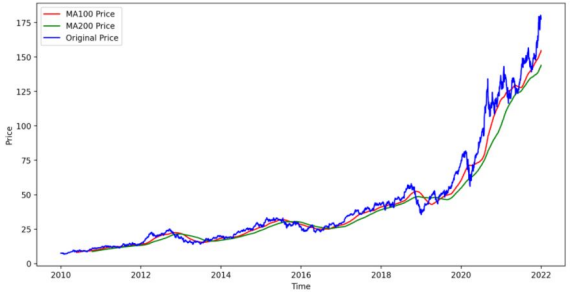
Data from 2010 - 2021

	High	Low	Open	Close	Volume	Adj Close
count	3,022.0000	3,022.0000	3,022.0000	3,022.0000	3,022.0000	3,022.0000
mean	43.1309	42.2492	42.6864	42.7091	270,286,158.2065	40.8660
std	38.4973	37.6207	38.0470	38.0835	225,919,269.8782	38.4925
min	7.0000	6.7946	6.8704	6.8589	41,000,000.0000	5.8645
25%	18.6494	18.3164	18.5072	18.5364	109,652,500.0000	16.1839
50%	28.5500	27.9800	28.2638	28.2625	184,235,200.0000	26.0292
75%	48.5988	47.7812	48.0725	48.1750	370,730,400.0000	46.5184
max	182.1300	178.5300	181.1200	180.3300	1,880,998,000.0000	179.8363

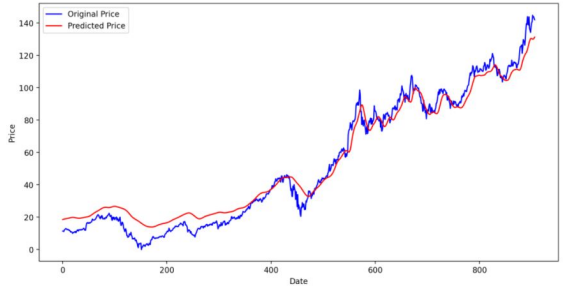
Closing price vs Time chart



Closing price vs Time chart with moving average

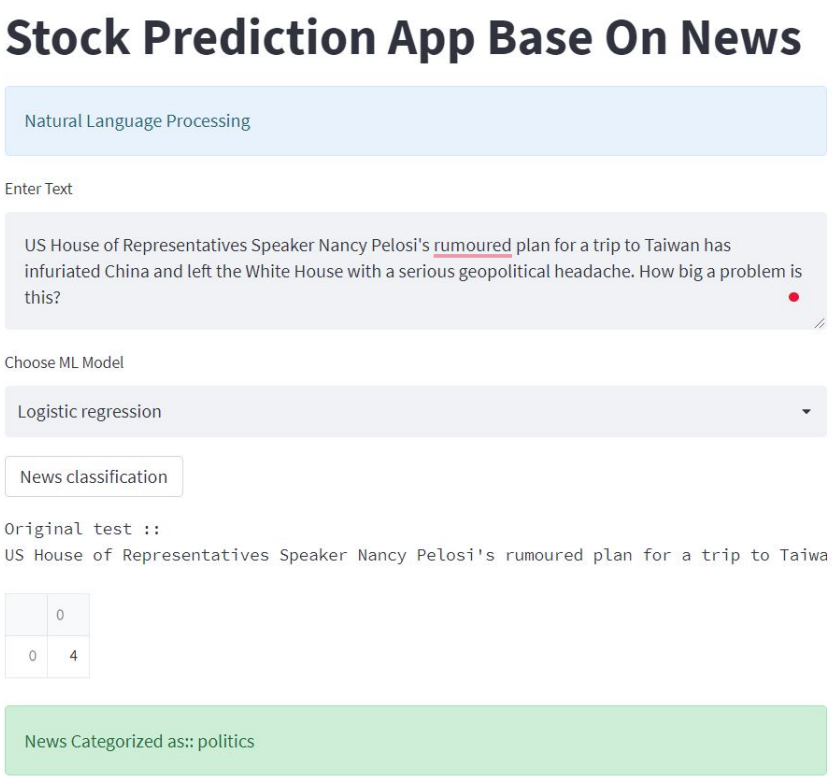


Predictions vs Original from 2019 to 2022



News classification based on different ML model

The app can classify the news based Logistic Regression, Random Forest and Decision Tree



News classification based on different ML model

The app can impelement different NLP tasks and generate a wordcould based on the news content

Choose NLP Task

Lemmatization

Analyze

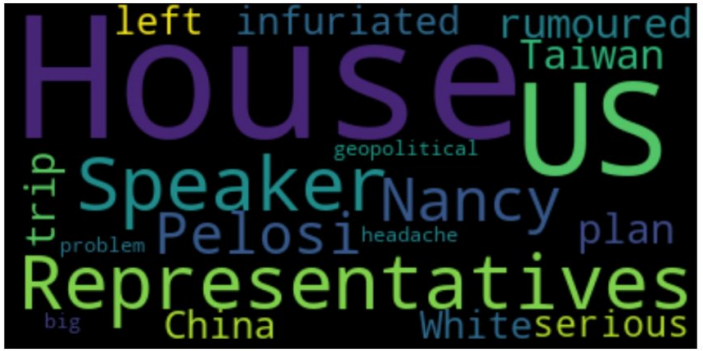
Original Text US House of Representatives Speaker Nancy Pelosi's rumoured plan for a trip to Taiwan has infuriated China and left the White House with a serious geopolitical headache. How big a problem is this?


```
[
0 : "'Token':US,'Lemma':US"
1 : "'Token':House,'Lemma':House"
2 : "'Token':of,'Lemma':of"
3 : "'Token':Representatives,'Lemma':Representatives"
4 : "'Token':Speaker,'Lemma':Speaker"
5 : "'Token':Nancy,'Lemma':Nancy"
6 : "'Token':Pelosi,'Lemma':Pelosi"
7 : "'Token':'s,'Lemma':'s"
8 : "'Token':rumoured,'Lemma':rumour"
9 : "'Token':plan,'Lemma':plan"
10 : "'Token':for,'Lemma':for"
11 : "'Token':a,'Lemma':a"
12 : "'Token':trip,'Lemma':trip"
13 : "'Token':to,'Lemma':to"
```


Tabulize

	Tokens	Lemma	POS
0	US	US	NNP
1	House	House	NNP
2	of	of	IN
3	Representa	Represen	NNPS
4	Speaker	Speaker	NNP
5	Nancy	Nancy	NNP
6	Pelosi	Pelosi	NNP
7	's	's	POS
8	rumoured	rumour	VBN
9	plan	plan	NN

Wordcloud



FUTURE

More Training

More Test

More Data

We can get:

The range of ups and downs of any stock related to news