

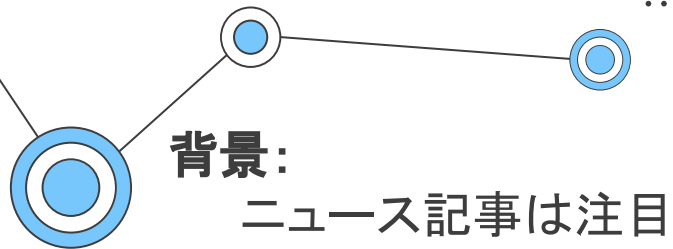
ニュースのキーワードによる 株価の時系列分析

名前: KHUN LUNN HTET

目次

- 背景と目的
- Bertについて
 - 事前学習
 - ファイルチューニング
 - Transformer
- 感情分析と可視化
- ARIMA分析
- 相関確認
- 学習
- 参考文献





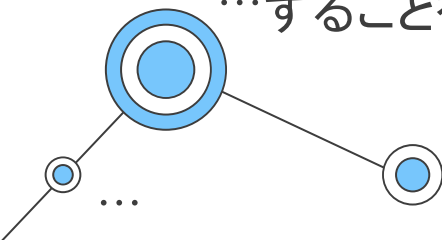
背景:

ニュース記事は注目を集めているため、多くのマスメディアが報道し、話題性が高くなり投資家はそれぞれの会社の株価を売買することにより株価が変動する。

金融市場とテキストデータ(ニュース、ブログ、SNS など)の人工知能の進歩により、一定の精度で株価の予測ができると考えられます。

目的:

本研究では、ニュースの記事と、その後の株価の関係性について分析
...することを目的とする。



BERT:

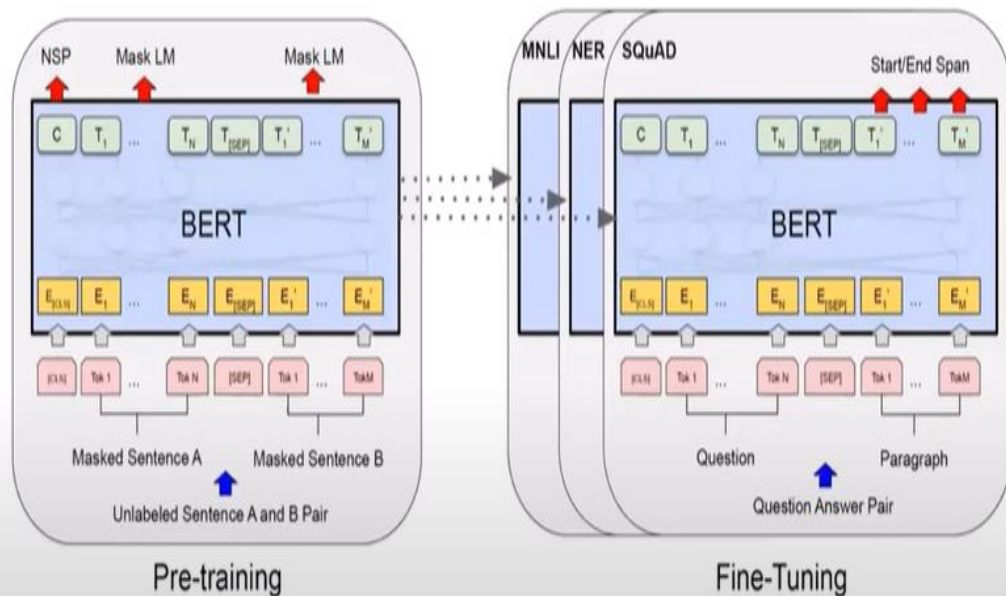
大量の文章のデータを用いて汎用的な言語のパターンを学習します。

大量のデータの事前学習によって、データを適切に表現するような「良い」特徴を抽出できるようになっている。

ファインチューニングによりBERTが抽出する特徴も問題に合わせて調整さる。

Transformer 様々な言語の事前学習モデルが利用可能

Pre-training , Fine-Tuning



データ

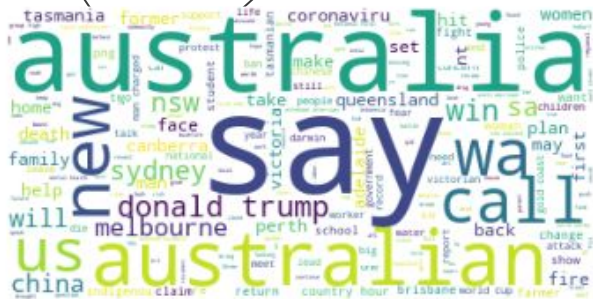
	Date	Open	High	Low	Close
0	2015-01-05	5437.1	5470.4	5429.6	5450.3
1	2015-01-06	5423.8	5424.3	5338.6	5364.8
2	2015-01-07	5362.2	5369.8	5318.7	5353.6
3	2015-01-08	5361.9	5395.8	5359.9	5381.5
4	2015-01-09	5393.2	5465.6	5393.2	5465.6
5	2015-01-12	5451	5451.5	5419.4	5422.7
6	2015-01-13	5411	5416.6	5373.7	5404.7
7	2015-01-14	5403.3	5405.7	5346.6	5353.6
8	2015-01-15	5347.5	5347.5	5303.7	5331.4
9	2015-01-16	5332.8	5333.9	5267.4	5299.2
10	2015-01-19	5321.5	5372.4	5302.3	5309.1
11	2015-01-20	5309.3	5313.9	5282.4	5307.7
12	2015-01-21	5314.8	5393.4	5314.8	5393.4
13	2015-01-22	5405.4	5432.4	5404.2	5419.9
14	2015-01-23	5436.3	5501.8	5436.3	5501.8

S&P/ASX200[1]

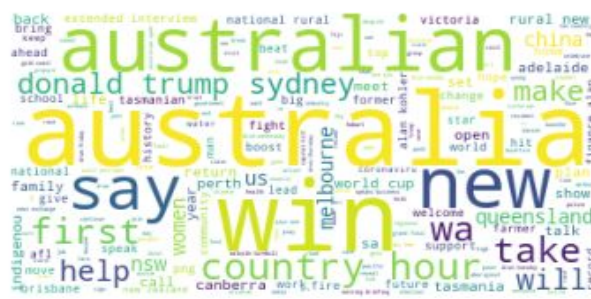
publish_date	headline_text
20030219	aba decides against community broadcasting licen
20030219	act fire witnesses must be aware of defamation
20030219	a g calls for infrastructure protection summit
20030219	air nz staff in aust strike for pay rise
20030219	air nz strike to affect australian travellers
20030219	ambitious olsson wins triple jump
20030219	antic delighted with record breaking barca
20030219	aussie qualifier stosur wastes four memphis match
20030219	aust addresses un security council over iraq
20030219	australia is locked into war timetable opp
20030219	australia to contribute 10 million in aid to iraq
20030219	barca take record as robson celebrates birthday in
20030219	bathhouse plans move ahead
20030219	big hopes for launceston cycling championship
20030219	big plan to boost paroo water supplies
20030219	blizzard buries united states in bills
20030219	brigadier dismisses reports troops harassed in
20030219	british combat troops arriving daily in kuwait
20030219	bryant leads lakers to double overtime win
20030219	bushfire victims urged to see centrelink
20030219	businesses should prepare for terrorist attacks
20030219	calleri avenges final defeat to eliminate massu
20030219	call for ethanol blend fuel to go ahead
20030219	carews freak goal leaves roma in ruins
20030219	cemeteries miss out on funds
20030219	code of conduct toughens organ donation regulatio

ABC[2]

感情分析と可視化



全体的の WordCloud



ポジティブの WordCloud



ネガティブの WordCloud

publish_date	label	score	codi_POSITIVE	codi_NEGATIVE	NEG_SCORE	POS_SCORE	
2015-01-01		63	122.7581524	63	67	62.9305374	59.82761496
2015-01-02		73	141.8737362	73	75	72.17736143	69.69637477
2015-01-03		25	55.1741206	25	33	31.43004644	23.74407417
2015-01-04		28	70.3039254	28	46	43.42580426	26.87812114
2015-01-05		32	79.22254997	32	52	49.31237286	29.91017711
2015-01-06		8	20.04368776	8	13	12.34954029	7.694147468
2015-01-07		9	24.60571897	9	17	15.95870107	8.647017896
2015-01-08		15	21.49503326	15	8	7.080055892	14.41497737
2015-01-09		13	20.13122863	13	8	7.800115645	12.33111298
2015-01-10		12	18.26044691	12	7	6.600646138	11.65980077
2015-01-11		11	18.61711776	11	8	7.785994232	10.83112353
2015-01-12		10	23.44308615	10	15	14.67608511	8.767001033
2015-01-13		19	48.25846726	19	33	30.55951548	17.69895178
2015-01-14		26	60.51447403	26	38	36.4247421	24.08973193
2015-01-15		25	54.73731166	25	33	31.31583488	23.42147678

感情分析したスコア



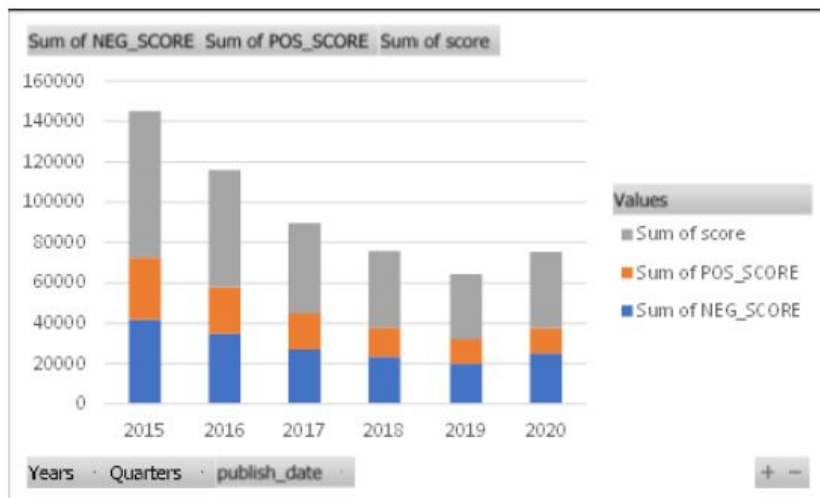


図 10. 年毎の Positive/Negative のスコアのヒストグラム

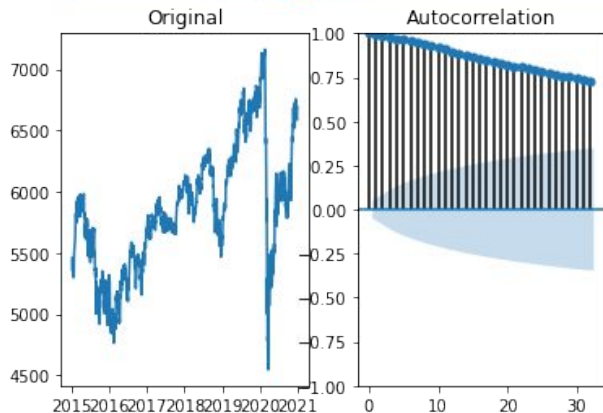
表 1. 年毎の Positive/Negative のスコアの内訳および和

Row Labels	Positive	Negative	Sum of score
2015	30804.09445	41832.95879	72637.05325
2016	23217.69357	34787.5785	58005.27207
2017	17932.28875	26884.33758	44816.62633
2018	14456.42962	23400.34047	37856.77009
2019	12089.11167	20059.00269	32148.11436
2020	12617.60123	25088.43563	37706.03686
Grand Total			283169.873

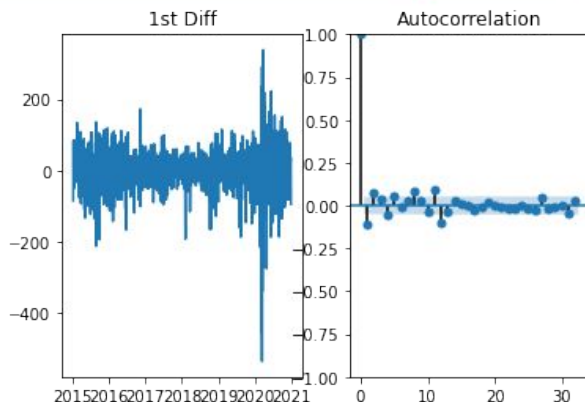
ARIMA分析



株価の移動平均



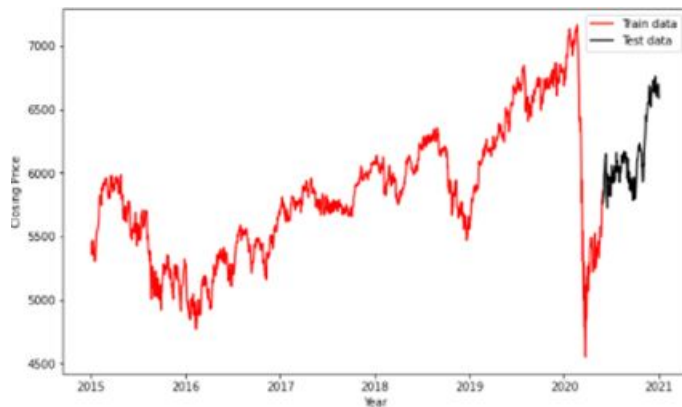
1. ADF : -2.342023405623188
2. P-Value : **0.15877243257602147**
3. Num Of Lags : 13
4. Num Of Observations Used For ADF Regression and Critical Values Calculation : 1504
5. Critical Values :
 - 1% : -3.4347053829922043
 - 5% : -2.863463625572207
 - 10% : -2.5677941141530387



1. ADF : -10.568935417124363
2. P-Value : **7.391254305515992e-19**
3. Num Of Lags : 12
4. Num Of Observations Used For ADF Regression and Critical Values Calculation : 1504
5. Critical Values :
 - 1% : -3.4347053829922043
 - 5% : -2.863463625572207
 - 10% : -2.5677941141530387

ARIMA分析

ARIMA分析

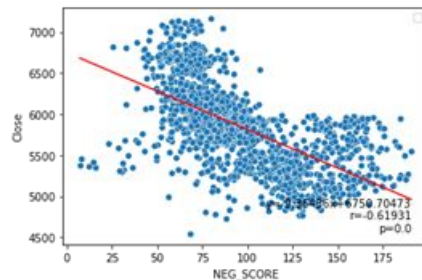


学習データとテスト
データ

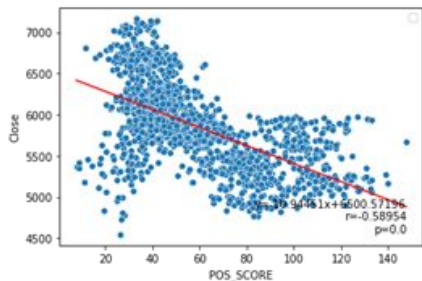


結果RMSE :410.547

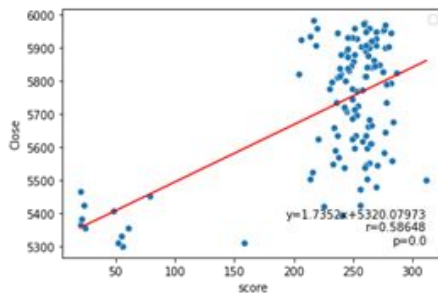
相関の確認



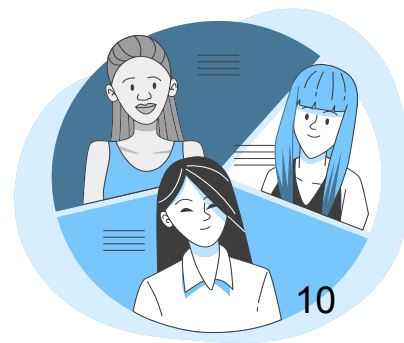
5年間の株とネガティブな感情スコアの相関(負)



5年間の株とポジティブな感情スコアの相関(負)

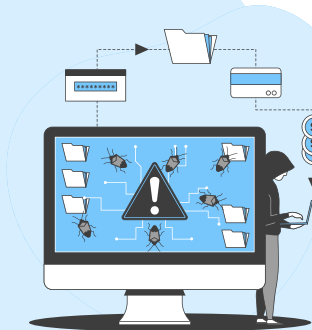
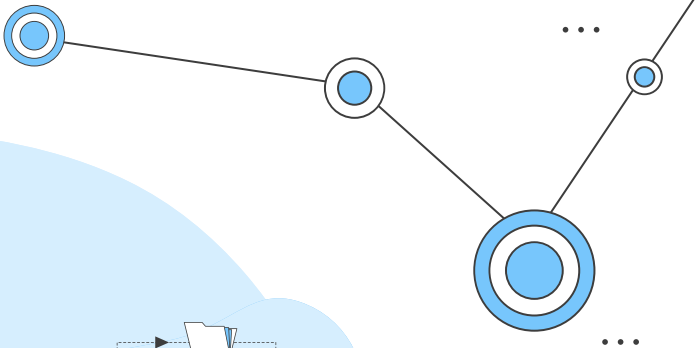


半年間の株と感情の相関(正)



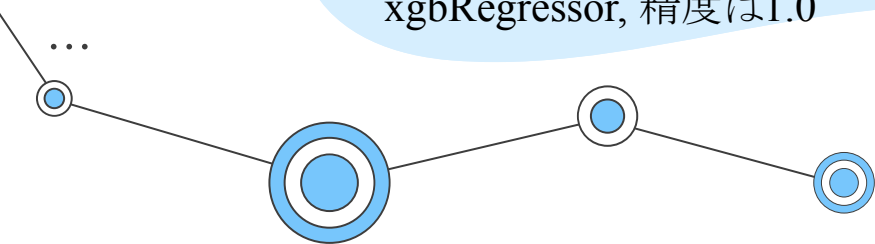
学習

Date	Score	Price	Diff	Trend
2015-01-06	20.043688	5364.8	-85.5	0
2015-01-07	24.605719	5353.6	-11.2	0
2015-01-08	21.495033	5381.5	27.9	1
2015-01-09	20.131229	5465.6	84.1	1
2015-01-12	23.443086	5422.7	-42.9	0



xgbRegressor, 精度は1.0

RandomForest, 精度は0.96710





Thanks!

ご清聴ありがとうございました

参考文献:

1. S&P/ASX200 (エスアンドピー/エーエスエックス) <https://www.smbcnikko.co.jp/>
2. Market Watch(株価のデータ) <https://www.marketwatch.com/>
3. スtockマーク株式会社編, 近江他著: BERT による自然言語処理入門, オーム社 (2021).
4. Hugging Face https://huggingface.co/docs/transformers/model_summary
5. WordCloud(WordCloud for Python documentation) https://amueller.github.io/word_cloud/

CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#), infographics & images by [Freepik](#) and illustrations by [Stories](#)