

# CP1

# Bitcoin Sentiment Analysis

AI10기 임규현  
5/2/21

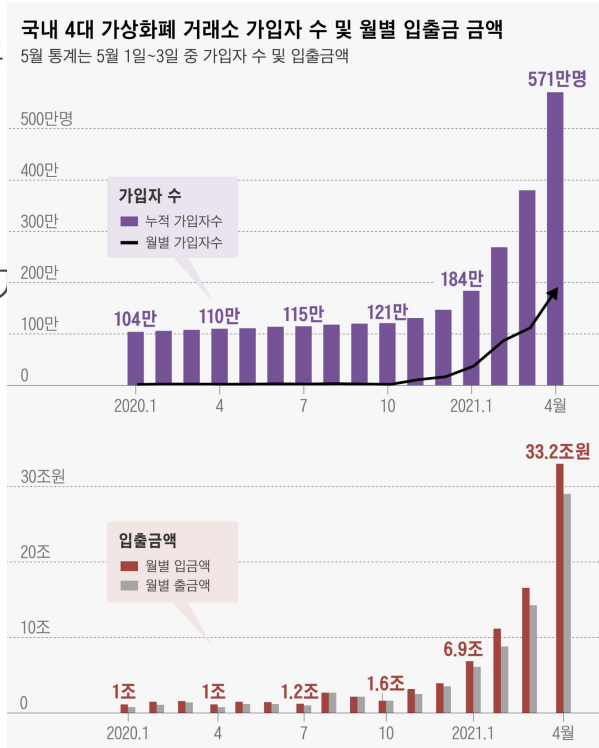


# 목차

- Background
- Problem Statement
- Data Pipeline
- Data Input & Processing
- Modeling
- Limitations / Improvements

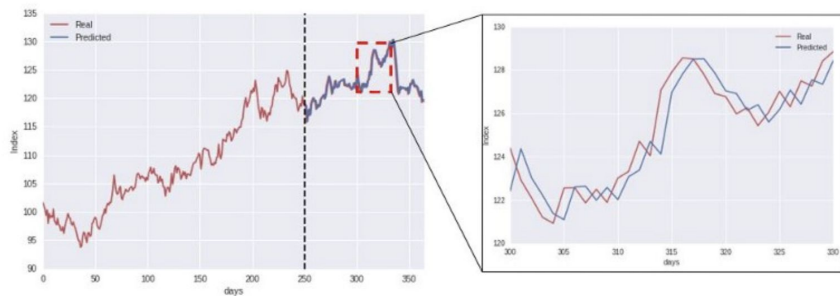
# Background

- 비트코인은 암호화 자산의 나스닥 / 코스피 같은
  - 암호화폐 시장의 40% 를 차지
- 늘어나는 기관 및 대중의 참여
- 딥러닝 모델이 비트코인 예측에 적용될 수 있을지
- 매매에 대한 기준?
  - 기분, 뉴스, 지인, 차트, 인간지표
- 인간지표의 숫자화?



# Problem Statement - 1

1. 일반적으로 온라인에서 접할 수 있는 가격 예측 관련 딥러닝 모델들의 문제



- 1.1. 돈복사?
- 1.2. Overfitting? 왜? -> Loss 값의 최적화 때문
  - 1.2.1. 최적화- 어제의 값이 내일의 값과 같다
- 1.3. 제대로 된 딥러닝 모델을 만드는 게 가능할까?

## Problem Statement - 2

### 1. 대안?

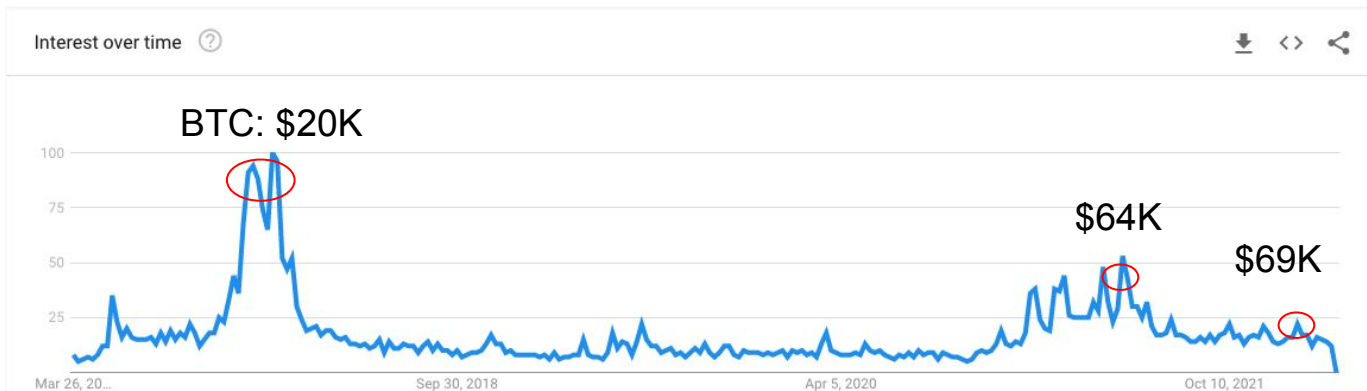
1.1. Non-stationary -> Stationary Data로의 전환 -> 모델링

1.2. 내일의 가격이 **상방**일지 **하방**일지 예측 (이중분류)

1.3. Feature?

1.3.1. Google Trend, 뉴스, 차트, 유동성, Twitter Sentiment Analysis

(+, -)



# Problem Statement - 2

## 1. 대안?

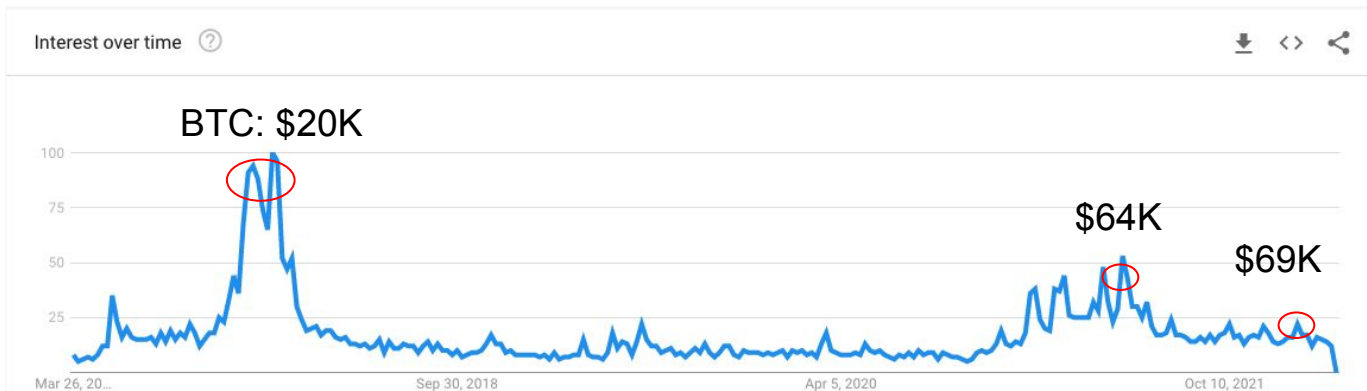
1.1. Non-stationary -> Stationary Data로의 전환 -> 모델링

1.2. 내일의 가격이 **상방일지 하방일지** 예측 (이중분류)

1.3. Feature?

1.3.1. Google Trend, 뉴스, 차트, 유동성, **Twitter Sentiment Analysis**

**(+, -)**





# Data Pipeline

데이터  
수집

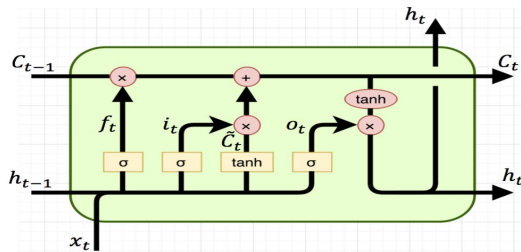
데이터  
전처리

모델링

모델 분석  
및 검증

서비스  
개발

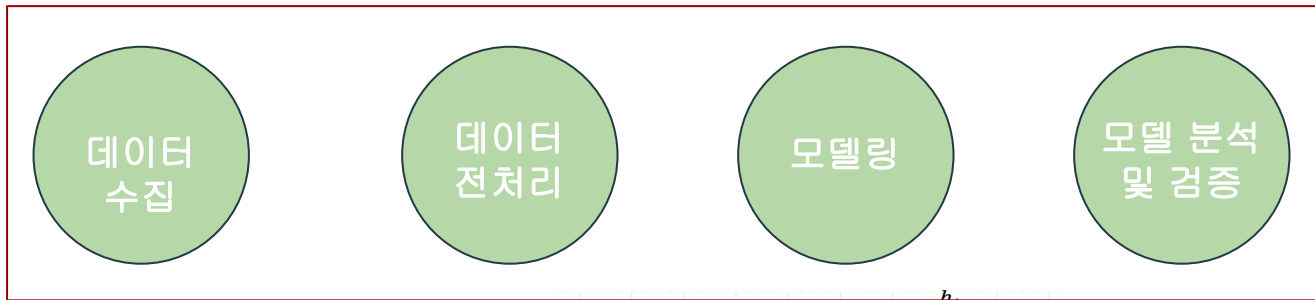
kaggle



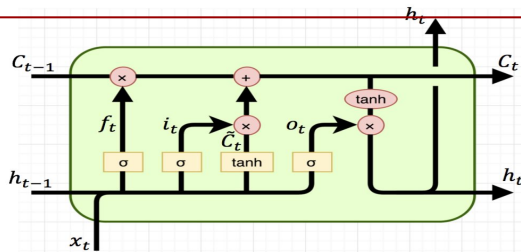
Dash  
byplotly



# Data Pipeline



kaggle



(a) Long Short-Term Memory



Dash  
byplotly





# Data Input & Processing - 1

- Kaggle의 'Bitcoin Tweets' 데이터 사용
  - 30만개의 트윗 \* 트윗에 대한 정보가 있는 13개의 열
- Alpha Vantage 의 API를 사용
  - 일별 비트코인 가격 데이터를 받아옴 (시가,저가,종가,거래량)
- 기간 설정
  - 2021년 2월 ~ 2022년 4월 -> 약 10,000 개의 트위터 데이터로 축소
- 텍스트 토큰화 후 단어 분포 분석
- VADER를 사용한 Sentiment Analysis 적용



# Data Input & Processing - 2

user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	
Hoba Bot	NaN	Fearless like a Honey Badger... Bears or Bulls...	2012-12-30 21:55:51	13144.0	7.0	143.0	False	2021-07-03 10:56:56	🚀🚀🚀 \$OI
defitralized • pøzīlē #ijəp	NaN	Defitralized © • DC of @_dcdao ...	2015-09-01 23:34:35	132.0	2387.0	8030.0	False	2022-03-15 09:40:37	discount\http
GORE I.A.D	Deutschland	telegram https://t.co/ui9meh3B41\nMOD https://...	2021-12-03 13:05:19	44.0	397.0	5689.0	False	2022-02-09 18:35:47	@crypto.
satsleft.info	The Bitcoin Timechain	Reporting on the remaining #Bitcoin supply to ...	2021-10-14 21:31:42	92.0	251.0	12.0	False	2021-11-11 17:17:32	👉 A new
Dweep	NaN	Vocalist, Musician, Author\n- CRYPTO maniac-	2017-08-17 15:01:18	8.0	49.0	283.0	False	2021-11-25 16:01:31	@CryptoWh
...	...	...	...	...	...	...	...	...	...
Bitcoin tips	NaN	NaN	2017-09-21 09:19:00	58.0	223.0	85.0	False	2021-12-17 13:26:16	@saylc
Somnath Sahoo 🟦	NaN	@Cardano @CotiNetwork \n@Ripple	2017-01-15 02:33:18	135.0	319.0	13924.0	False	2021-08-24 16:55:50	#Bitcoin N

# Data Input & Processing - 3

- 텍스트 토큰화 후 단어 분포 분석
  - 단어 토큰화를 통해 비트코인 관련 트윗에는 어떤 내용들이 많이 언급되는지 파악
  - 불용어처리를 통해 반복적으로 비슷한 단어들을 제거 (e.g. btc, \$btc, bitcoins)

	word	word_in_docs	count	rank	percent	cul_percent	word_in_docs_percent
1	#bitcoin	5703	5775	2.0	0.047656	0.101551	0.5703
36	#crypto	1245	1249	3.0	0.010307	0.111858	0.1245
151	buy	520	583	4.0	0.004811	0.116669	0.0520
142	price	557	569	5.0	0.004695	0.121365	0.0557
372	@elonmusk	480	492	6.0	0.004060	0.125425	0.0480
95	#eth	481	487	7.0	0.004019	0.129444	0.0481
634	follow	438	440	8.0	0.003631	0.133075	0.0438
575	new	390	402	9.0	0.003317	0.136392	0.0390
480	market	394	398	10.0	0.003284	0.139677	0.0394
72	prices	372	372	11.0	0.003070	0.142746	0.0372
482	best	227	345	12.0	0.002847	0.145593	0.0227
287	#dogecoin	325	341	13.0	0.002814	0.148407	0.0325
256	platform	337	338	14.0	0.002789	0.151197	0.0337
249	let's	334	336	15.0	0.002773	0.153969	0.0334
259	#1	334	334	16.0	0.002756	0.156726	0.0334
954	time	323	334	17.0	0.002756	0.159482	0.0323
376	#tesla	324	325	18.0	0.002682	0.162164	0.0324
493	24	205	313	19.0	0.002583	0.164747	0.0205
731	experience	297	297	20.0	0.002451	0.167198	0.0297
80	\$xrp	292	292	21.0	0.002410	0.169607	0.0292
733	exciting	290	290	22.0	0.002393	0.172000	0.0290
735	wi...	288	288	23.0	0.002377	0.174377	0.0288
737	👉	287	287	24.0	0.002368	0.176745	0.0287
738	together!	286	286	25.0	0.002360	0.179105	0.0286
736	i-gaming	286	286	26.0	0.002360	0.181466	0.0286

# Data Input & Processing - 4

- VADER 를 통한 Sentiment Analysis
  - Natural Language Toolkit (NLTK)에 있는 감성 분석기
  - 단어들을 긍,부,중립으로 분류해줄 수 있음
  - 소셜미디어 분석을 위해 만들어진 분석기

표 5. 대화별 VADER 값 예시








작품	막 및 장 번호	발화자	대화	VADER 값
Othello	5막 2장	에밀리아	Evil, evil, evil! I can smell it! I suspected it earlier. I'll kill myself out of grief! Oh, evil, evil!	-0.987
King Lear	1막 1장	고너릴	Sir, I love you more than words can say. I love you more than eyesight, space, and freedom, beyond wealth or anything of value. I love you as much as life itself, and as much as status, health, beauty, or honor. I love you as much as any child has ever loved her father, with a love too deep to be spoken of. I love you more than any answer to the question "How much?"	0.993

딥러닝 모델 학습을 위해 전치리틀 한 문장들을 NLTK VADER 감정분석기에 동일하게 적용하여 각 대화의 VADER 감정값을 구했다.



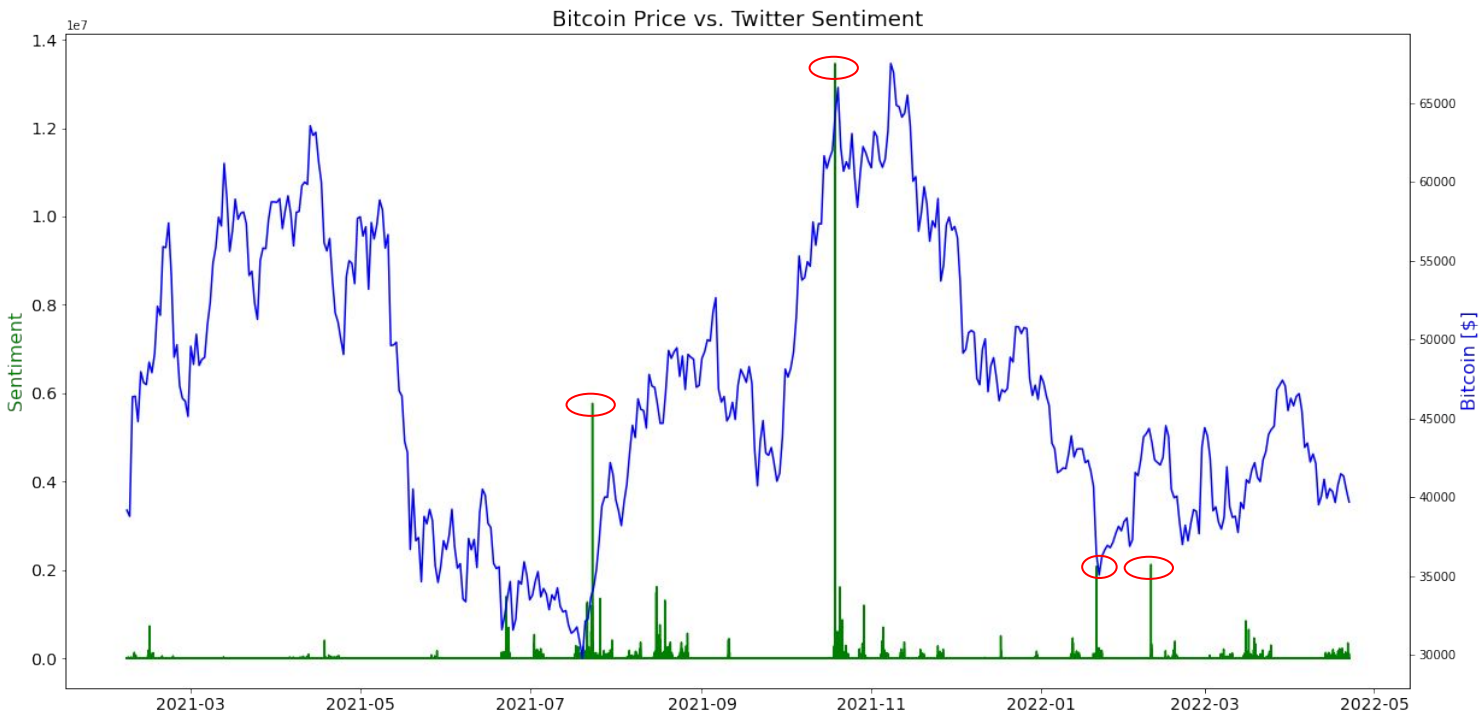
# Data Input & Processing - 5

- VADER 적용후 데이터

avourites	user_verified	date	text	...	source	is_retweet	cleantext	vader_neg	vader_neu	vader_pos	vader_comp	cleantext2	class
143.0	False	2021-07-03 10:56:56	   \$ONT Long Position(v2)    \n\nEntry: 0.717...	...	HoBaBot	False	['ont', 'long', 'posit', 'v2', 'entri', '0', '...	0.128	0.872	0.000	-0.2960	ont long posit v2 entri 0 7179 target 0 7213 s...	-1
8030.0	False	2022-03-15 09:40:37	discount\nhttps://t.co/1360Y2IRC7\n\n#BTC...	...	Twitter for Android	False	['10', 'discount', 'http', 'co', '1360y2irc7', ...	0.000	1.000	0.000	0.0000	10 discount http co 1360y2irc7 btc inliqwetrus...	0
5689.0	False	2022-02-09 18:35:47	@crypto_bearr Join us In #Aether Before the la...	...	Twitter for Android	False	['crypto', 'bearr', 'join', 'us', 'aether', 'l...	0.000	0.936	0.064	0.2960	crypto bearr join us aether launch satellit or...	1
12.0	False	2021-11-11 17:17:32	 A new block was found on the #Bitcoin networ...	...	satsleft	False	['new', 'block', 'found', 'bitcoin', 'network ...	0.293	0.707	0.000	-0.7003	new block found bitcoin network block height 7...	-1
283.0	False	2021-11-25 16:01:31	@CryptoWhale Its just whale games. Plan B made...	...	Twitter for Android	False	['cryptowhal', 'whale', 'game', 'plan', 'b', '...	0.000	0.791	0.209	0.7003	cryptowhal whale game plan b made whole market...	1

# Data Input & Processing - 6

- 비트코인 가격 vs. 트위터 반응





# Modeling - 1

- Sentiment 예측 모델
  - a. 다중분류 문제로 접근 (Activation= Softmax, Loss='categorical\_crossentropy')
    - i. 데이터 레이블링 - 긍정:1, 중립:0, 부정: -1
  - b. Sentiment 계수와 가격과의 상관관계가 성립이 된다면 결과의 신뢰도가 높을 것으로 예상
  - c. LSTM

```
Epoch 95/100
38/38 [=====] - 11s 278ms/step - loss: 0.1170 - accuracy: 0.9710 - val_loss: 0.6943 - val_accuracy: 0.8375
Epoch 96/100
38/38 [=====] - 10s 277ms/step - loss: 0.1147 - accuracy: 0.9710 - val_loss: 0.7332 - val_accuracy: 0.8425
Epoch 97/100
38/38 [=====] - 11s 277ms/step - loss: 0.1116 - accuracy: 0.9715 - val_loss: 0.6745 - val_accuracy: 0.8458
Epoch 98/100
38/38 [=====] - 11s 281ms/step - loss: 0.0952 - accuracy: 0.9737 - val_loss: 0.5935 - val_accuracy: 0.8500
Epoch 99/100
38/38 [=====] - 11s 283ms/step - loss: 0.0787 - accuracy: 0.9748 - val_loss: 0.6532 - val_accuracy: 0.8550
Epoch 100/100
38/38 [=====] - 11s 281ms/step - loss: 0.0918 - accuracy: 0.9613 - val_loss: 0.6842 - val_accuracy: 0.8642
<keras.callbacks.History at 0x7f2ed830afd0>
```

---

[0.7112246751785278, 0.8690000176429749]

---

# Modeling - 2

- Sentiment Analysis Score를 통한 비트코인 방향성 예측
  - LSTM 사용
  - 이중분류 문제로 접근 (Activation= Sigmoid, Loss='binary\_crossentropy')

```
model2 = tf.keras.models.Sequential([
    tf.keras.layers.Embedding(10000, 128),
    tf.keras.layers.LSTM(128),
    tf.keras.layers.Dense(1, activation='sigmoid')])

model2.compile(loss='binary_crossentropy',
               optimizer='adam',
               metrics=['accuracy'])
```

] # 세부적인 튜닝 필요함

```
model2.fit(X_train, Y_train,
          batch_size=128,
          epochs=100,
          validation_split = 0.2)
```

```
1/1 [=====] - 0s 146ms/step - loss: 0.6855 - accuracy: 0.5056 - val_loss: 0.6935 - val_accuracy: 0.4783
Epoch 95/100
1/1 [=====] - 0s 116ms/step - loss: 0.6854 - accuracy: 0.5056 - val_loss: 0.6930 - val_accuracy: 0.5217
Epoch 96/100
1/1 [=====] - 0s 131ms/step - loss: 0.6854 - accuracy: 0.5056 - val_loss: 0.6932 - val_accuracy: 0.4783
Epoch 97/100
1/1 [=====] - 0s 148ms/step - loss: 0.6854 - accuracy: 0.5056 - val_loss: 0.6933 - val_accuracy: 0.4783
Epoch 98/100
1/1 [=====] - 0s 145ms/step - loss: 0.6854 - accuracy: 0.5056 - val_loss: 0.6929 - val_accuracy: 0.5217
Epoch 99/100
1/1 [=====] - 0s 129ms/step - loss: 0.6854 - accuracy: 0.5056 - val_loss: 0.6935 - val_accuracy: 0.4783
Epoch 100/100
1/1 [=====] - 0s 131ms/step - loss: 0.6854 - accuracy: 0.5056 - val_loss: 0.6928 - val_accuracy: 0.5217
<keras.callbacks.History at 0x7f2ee340cb50>
```

[0.6931901574134827, 0.5]





## Modeling - 3

- DL을 통해 시장 참여자들의 Sentiment(감성) 예측은 꽤나 높은 정확도를 보인다
- 현재 모델을 기준으로 봤을 때 Sentiment의 비트코인에 대한 예측력은 높지 않은 수준
  - 사실상 훌쩍 혹은 찍기
- 하지만 Sentiment vs. Price 그래프를 보았을 때
  - 긍정적이든 부정적이든 많은 Sentiment는 높은 가격 변동성으로 이어짐
  - 변동성에 익숙하지 않은 사람들은 레버리지를 줄이는 게 멘탈관리에 좋을 것으로 판단



## Limitations / 아쉬운 점

- 하고자 하는 건 많았으나 구현하는데 어려움이 있었음 (서비스 개발 - DASH)
- DL 모델 성능 개선
  - 하이퍼파라미터 튜닝
- Sentiment ~ 가격/가격 방향 관계에 대한 수치화

## Improvements

- Vader 외 다른 Sentiment Analysis 도구?
- Sentiment 기반 가격 예측 모델
- 종합적인 DL 모델 만들기
- 트위터 외(뉴스 기사 등) Sentiment 분석