Analysing Twitter Data to Predict the Price of Stock Market Indices

Atai Otorbaev

August 2020

Background

Exploratory study;

The aim to investigate if sentiment expressed on Twitter has an effect on Price of Individual Stock Market Indices.

- FTSE

- Can Twitter Data be used to predict the Price of an Index up to one day ahead?
- Is Twitter a suitable data source to help forecast future stock market movement ?

Hypothesis

1. Twitter sentiment reflects the mood of the market.
2. Twitter can be used to predict future stock market movements.

Process of Getting the Data Frame

Scraping Twitter

Scraping tweets that mention the Index.

-Date, From: 2007.02.01

-^for a full economic cycle

- 2,497,676 tweets(FTSE)

Assigning Score

Assigning Sentiment Score to each Tweet

-VADER

- Loughran-McDonald finance word dictionary

- Assigned scores to the words in the dictionary

-Makes scores more **Precise**

Cleaning Data

Removing tweets with a null sentiment score

- 1,360,736 tweets(FTSE)

Bucket Tweets

Bucket real time tweets into bindates.

-3293 rows(FTSE)

-Stock exchanges close at 4.00pm. So assign tweets tweeted after to next working day.

-Applies to weekends tweets, will be assigned to Monday.

Consolidating Data

Computing the mean and weighted mean(favourites, retweets,replies) score of all tweets for specific days.

Index historical data

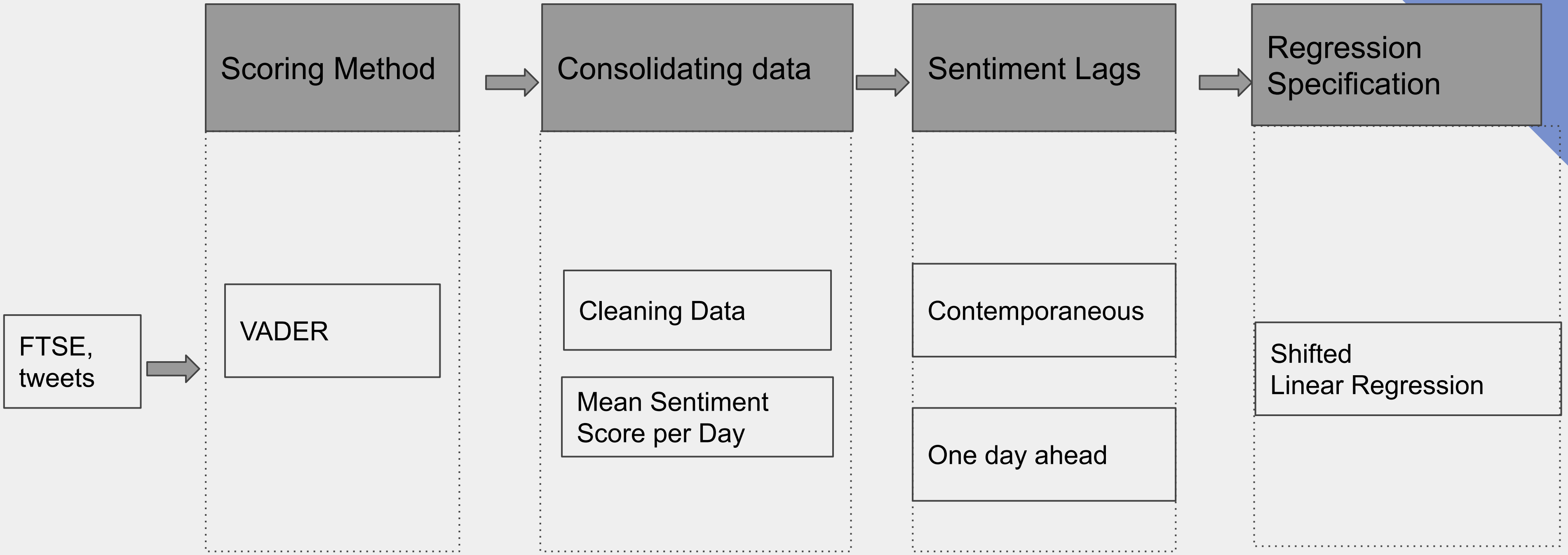
Loading the Indices historical data, and calculating Returns.

-Percentage change of the Closing price

- Right join

-3533 rows(FTSE)

Steps of Collecting and processing Data



Individual Tweets and Sentiment Scores

2010-03-12	240	2010-03-12 11:07:34+00:00	British Airways shares up on FTSE 100 as strike seems imminent http://bit.ly/apfrlt	IBTimesUK		0	0		0	47916714	0.3612	11	2010-03-12
2010-03-12	241	2010-03-12 11:07:31+00:00	Luminar shares down on FTSE Small Cap after sales fall 10% http://bit.ly/aO9Y0b	IBTimesUK		0	0		0	47916714	0.0516	11	2010-03-12
2010-03-12	242	2010-03-12 11:05:09+00:00	New Review: FTSE latest: RBS, M&S up; Luminar down http://bit.ly/bV22Z6	redbuttonreview		0	0		0	72275306	-0.25	11	2010-03-12
2010-03-12	243	2010-03-12 10:56:26+00:00	FTSE Stocks: Happy to be long lots of things, but worried that the general quietness of the markets suggests we're about to turn over...	FuturesTechs		0	0		0	54536679	-0.3818	10	2010-03-12
2010-03-12	244	2010-03-12 10:40:16+00:00	Market Update: Commods, financials push FTSE down 0.4%; China data: Extract not available. http://bit.ly/diazEQ www.stock-trkr.co.uk	StockTrkr		0	0		0	69046380	-0.25	10	2010-03-12
2010-03-12	245	2010-03-12 10:38:12+00:00	FTSE latest: RBS, M&S up; Luminar down: Continued uncertainty about the global economic outlook ensured the Londo... http://bit.ly/bHeDer	Market_Tweet		0	0		0	48563792	-0.296	10	2010-03-12
2010-03-12	246	2010-03-12 10:36:49+00:00	sensex: FTSE flat; weak oils, miners balance firm banks: Oils, miners slip back on China tightening uncertainties... http://bit.ly/cl9Anp	sensextweet		0	0		0	18483922	-0.4939	10	2010-03-12
2010-03-12	247	2010-03-12 10:35:22+00:00	FTSE 100 Accelerated Returns Issue 5 – 3 days left to go! http://bit.ly/cG8RWH	BarclaysInvest		0	0		0	52342885	0	10	2010-03-12
2010-03-12	248	2010-03-12 10:35:16+00:00	#mkt U.K. FTSE 100 index up 0.4% at 5,641.02: This is a Real-time headline. These are breaking news, delivered http://url4.eu/1IGGt	eForexSystems	#mkt	0	0		0	112694268	0	10	2010-03-12
2010-03-12	249	2010-03-12 10:35:01+00:00	Financial stocks push FTSE 100 higher: London shares opened a little higher on Friday, capping off... http://bit.ly/aS699x #finance #money	finance_yard	#finance #	0	0		0	36262658	0.296	10	2010-03-12
2010-03-12	250	2010-03-12 10:34:58+00:00	FTSE 100 edges higher as financials outweigh mining falls: Despite mining groups weighing on the m... http://bit.ly/9vkjEg #finance #money	finance_yard	#finance #	0	0		0	36262658	0.25	10	2010-03-12
2010-03-12	251	2010-03-12 10:29:57+00:00	Group Financial Planning & Analysis Manager...: This leading FTSE Retail group, with a £multimillion turnover, boas ... http://bit.ly/bDvpYB	twittlatestjobs		0	0		0	108035735	0.3071	10	2010-03-12
2010-03-12	252	2010-03-12 10:29:37+00:00	FTSE flat; weak oils, miners balance firm banks http://bit.ly/cau1IT	Morebalanceinfo		0	0		0	80560300	-0.25	10	2010-03-12
2010-03-12	253	2010-03-12 10:24:56+00:00	Financial stocks push FTSE 100 higher http://bit.ly/bl8bGf	buffetfx		0	0		0	67287962	0	10	2010-03-12
2010-03-12	254	2010-03-12 10:22:52+00:00	FTSE 100 opens lower (AFP): AFP - Leading shares weakened on Thursday after unconvincing gains overnight on http://url4.eu/1IFpl	rapidsuccessnow		0	0		0	59218553	0.25	10	2010-03-12
2010-03-12	255	2010-03-12 10:21:08+00:00	US stock futures mixed ahead of retail sales data (Reuters - UK Focus) http://bit.ly/9UC5wQ	FTSE_Tweets		0	0		0	23181022	0	10	2010-03-12
2010-03-12	256	2010-03-12 10:16:54+00:00	BSkyB takeover talk helps lift FTSE 100 http://bit.ly/ahf5vK	Financial_Mind		0	0		0	55938111	0.3818	10	2010-03-12
2010-03-12	257	2010-03-12 10:16:53+00:00	FTSE 100 falters as China fears hit miners | World Latest News http://bit.ly/cqyo3v	Financial_Mind		0	0		0	55938111	-0.25	10	2010-03-12
2010-03-12	258	2010-03-12 10:16:15+00:00	Piazza Affari il lieve calo: Fonte: La Stampa Il Ftse Mib ha chiuso a -0,43%, il Ftse Italia All-Share a -0,38%,... http://bit.ly/dfUIIK	informazione		0	0		0	26788605	0.34	10	2010-03-12

bindates

text

username

Hashtags
Favorites,
Retweets,
Replies

Scores

Sentiment Analysis With VADER

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
pos_words = ['up','bull','high']
neg_words = ['down','bear','low']
uncertain_words = ['shaky']
words = pd.read_csv('finance_dict.csv')
for col in words.columns:
    words[col] = words[col].str.lower()
pos_words = pos_words + list(words.positive)
neg_words = neg_words + list(words.negative)
uncertain_words = uncertain_words + list(words.uncertain)

new_words_dict = {}
for word in pos_words + neg_words + uncertain_words:
    if word in pos_words:
        new_words_dict[word] = 1
    elif word in neg_words:
        new_words_dict[word] = -1
    else:
        new_words_dict[word] = -0.2

analyzer.lexicon.update(new_words_dict)
```

Loughran-McDonald finance word dictionary

	negative	positive	uncertain
0	abandon	able	abeyance
1	abandoned	abundance	abeyances
2	abandoning	abundant	almost
3	abandonment	acclaimed	alteration
4	abandonments	accomplish	alterations
...	...	...	...
2350	wrongdoing	NaN	NaN
2351	wrongdoings	NaN	NaN
2352	wrongful	NaN	NaN
2353	wrongfully	NaN	NaN
2354	wrongly	NaN	NaN

2355 rows × 3 columns

FTSE DataFrame

bindate	score	score_retweets	score_replies	score_favorites	retweets	favorites	replies	Close	Returns
2013-01-01	0.050727	-0.208900	0.320875	0.000000	18.0	0.0	4.0	5897.81	0.000000
2013-01-02	0.070267	0.174218	0.293008	0.095222	99.0	9.0	26.0	6027.37	0.021967
2013-01-03	0.211909	0.238970	0.120747	0.303846	94.0	13.0	34.0	6047.34	0.003313
2013-01-04	0.050287	0.122616	0.178419	0.478550	57.0	4.0	26.0	6089.84	0.007028
2013-01-07	0.178135	0.054998	0.154537	0.137562	130.0	26.0	19.0	6064.58	-0.004148
...	...	...	...	...	...	...	...	...	...
2020-08-10	0.146385	0.255829	0.276632	0.279631	242.0	693.0	97.0	6050.59	0.003052
2020-08-11	0.179729	0.426062	0.210567	0.382079	128.0	490.0	91.0	6154.34	0.017147
2020-08-12	0.116080	0.210003	0.188777	0.246521	298.0	936.0	129.0	6280.12	0.020438
2020-08-13	0.133928	0.181326	0.319378	0.307921	359.0	1347.0	156.0	6185.62	-0.015047
2020-08-14	0.236990	0.612014	0.606838	0.611605	505388.0	793472.0	7739.0	6090.04	-0.015452

1983 rows × 9 columns

Correlation between Scores and Returns

```
def plot_corr(s):
    s.reset_index().rename(columns={0:'corr'}).plot(kind='line', x='bindate', y='corr', figsize=(11,6))
for var in ['score', 'score_retweets', 'score_replies', 'score_favorites']:
    print(var)
    print(ftse[var].corr(ftse['Returns']))
    print(ftse[var].shift(1).corr(ftse['Returns']))
    s = ftse[var].rolling(window=22*12,min_periods=1).corr(ftse['Returns'])
    plot_corr(s)
    plt.title(var)
```

score
0.30324652366925187
-0.06164192448719644

score_retweets
0.1927819549318894
-0.012090713773427523

score_replies
0.1218602434673302
0.0019632194118197298

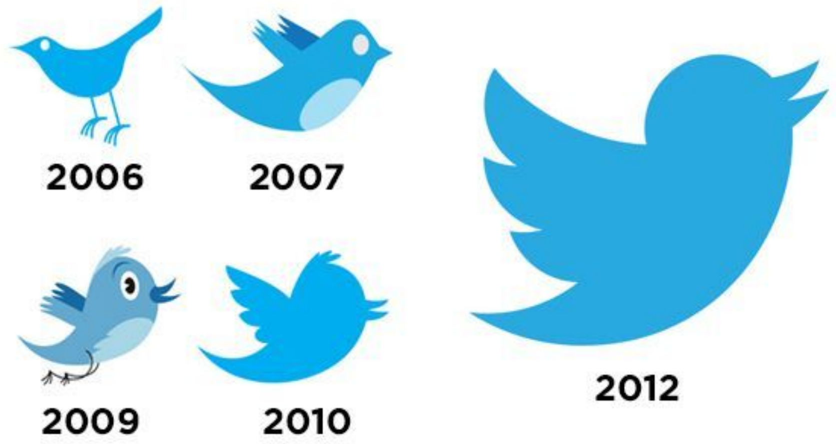
score_favorites
0.1230239624203053
0.007079182059807382

score-2013
0.32798691276405945
-0.015850594809868884

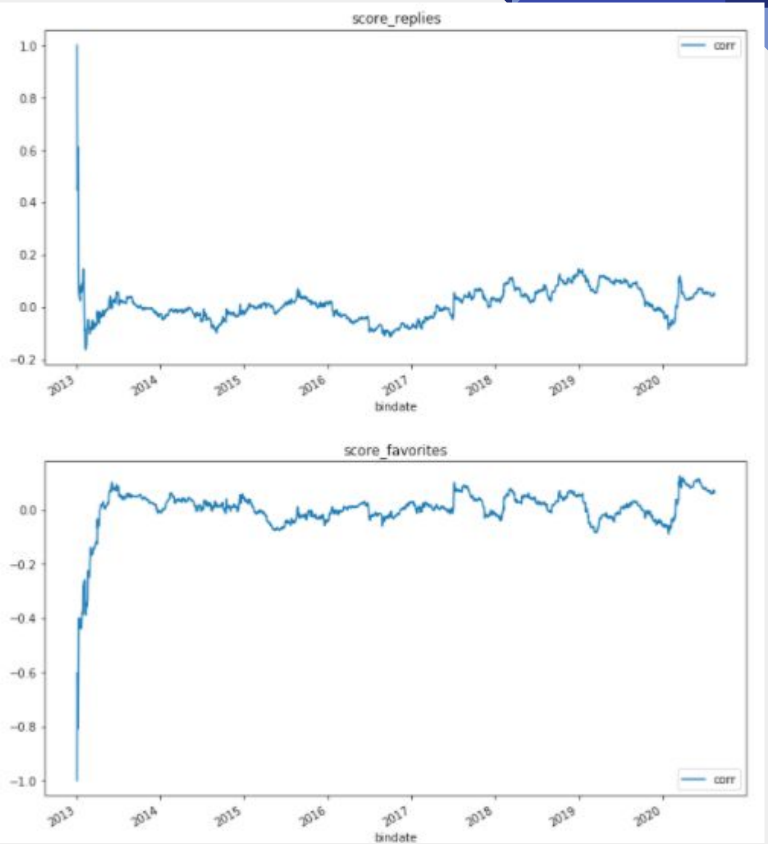
score_retweets-2013
0.20919808145477406
-0.005416047601653761

score_replies-2013
0.14483136453883919
0.016877667071874606

score_favorites-2013
0.19376169807880445
0.02048940286074888



EDA, Shifted Correlation between Scores and Returns since 2013 (one day ahead)



Shifted Linear Regression

Baseline: 0.028240040342914774

Model:

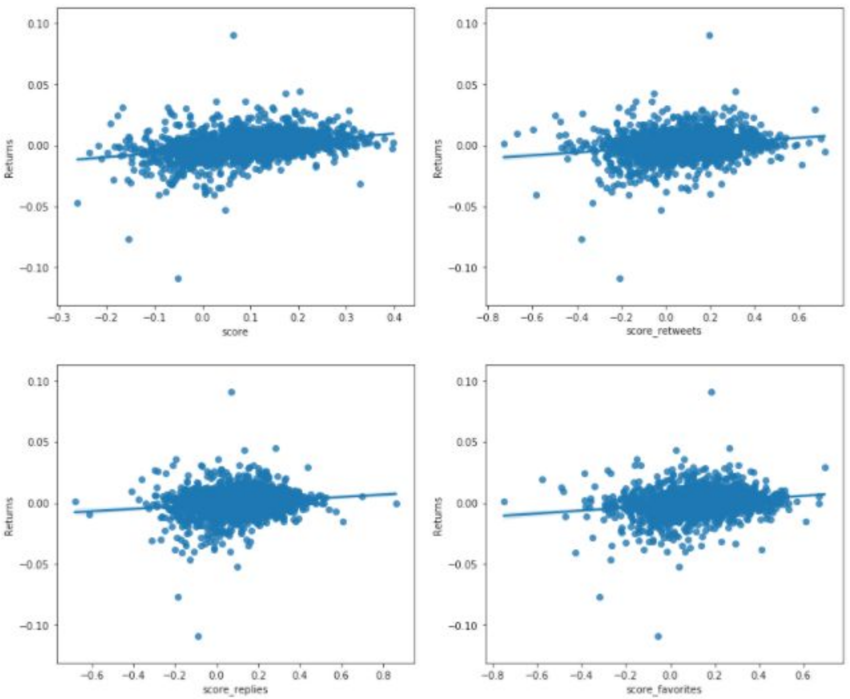
- Cross-validated training scores: [-0.0052319 -0.00363296 -0.00359711 -0.00285384 -0.01098714]
- Mean cross-validated training score: -0.0052605916793220684
- Training Score: 0.002574537329825488
- Test Score: 0.0009071482693060462

Not a good fit

Daily Volatility in the market!

From a Statistical standpoint it's a very bad model.

Scatterplots



But Economically it could be Rewarding!!

Formulate a Good Trading Strategy

Trading Strategy: A trading strategy is the method of buying and selling in markets that is based on predefined rules used to make trading decisions.

Backtesting assesses the viability of a trading strategy by discovering how it would play out using historical data. If backtesting works, traders and analysts may have the confidence to employ it going forward.

Evaluate Accumulated returns | P&L Vector

Make a Decision on which position to take based on Twitter sentiment at the closing time.

- Having a “**long**” position in a security means that you own the security. Investors maintain “**long” security positions in the expectation that the stock will rise** in value in the future. The opposite of a “long” position is a “short” position.

- A "**short**" position is generally the sale of a stock you do not own. Investors who **sell short believe the price of the stock will decrease in value**. If the price drops, you can buy the stock at the lower price and make a profit.

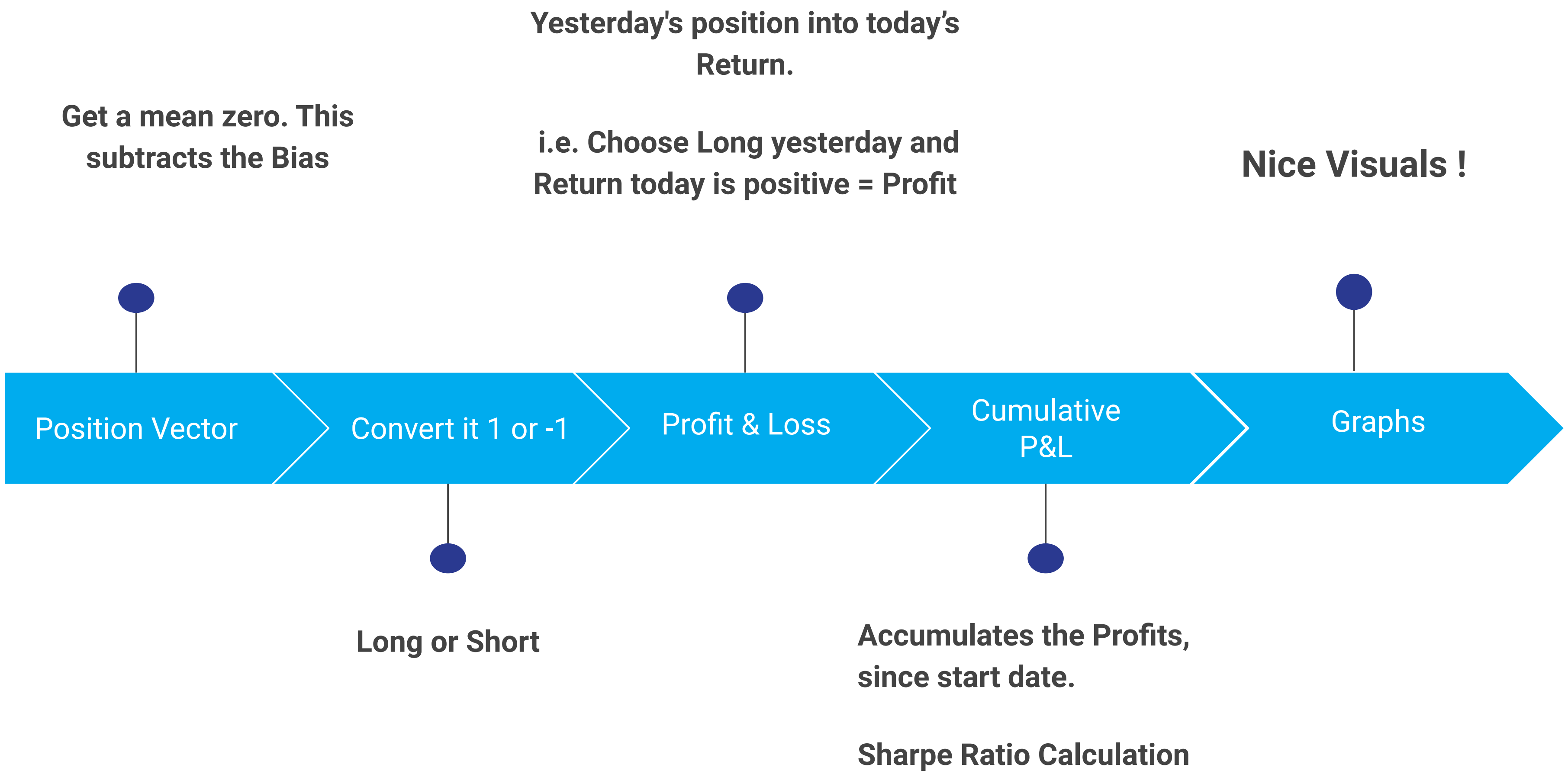
- **Trading Strategy:**

  Choose a position either **long** or **short** depending on if the twitter sentiment score on a specific date is greater or lower than the last weeks/ 5 business days (time period) average.

Trading Strategy Process

```
def plot_pnl(s):
    s = s.to_frame(name='Cumulative_P&L')
    s = s.reset_index()
    s.plot(kind='line', x='bindate', y='Cumulative_P&L', figsize=(11,6) )
for var in ['score', 'score_retweets', 'score_replies', 'score_favorites']:
    pos = ftse13[var]
    print(var)
    pos = pos - pos.rolling(window=5,min_periods=1).mean()  #subtracts bias
    # position vector long or short it
    pos = np.sign(pos)#convert to 1 or -1
    pnl = pos.shift(1) * ftse.Returns
    cumpnl = pnl.cumsum()
    plot_pnl(cumpnl)
    plt.title(var)
```

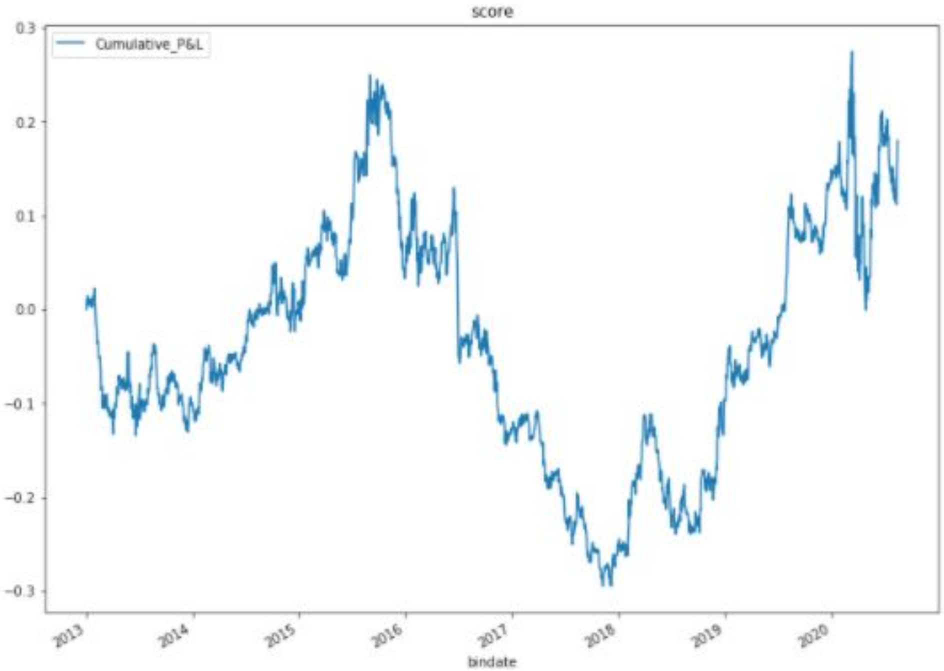
Position Vector → Convert it 1 or -1 → Profit & Loss → Cumulative P & L → Graphs

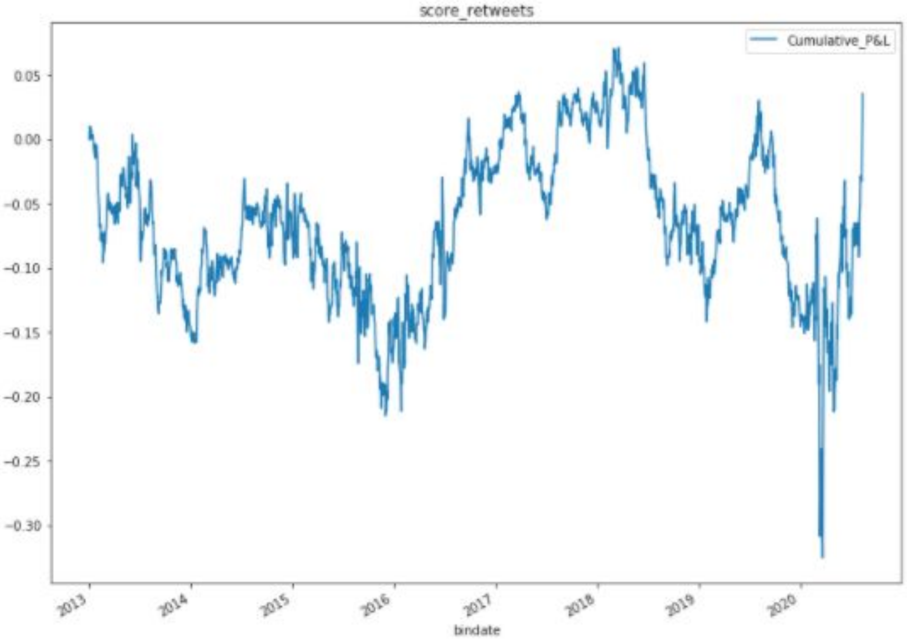

Yesterday's position into today's Return.
i.e. Choose Long yesterday and Return today is positive = Profit
Get a mean zero. This subtracts the Bias
Nice Visuals !
Position Vector
Convert it 1 or -1
Profit & Loss
Cumulative P&L
Graphs
Long or Short
Accumulates the Profits, since start date.
Sharpe Ratio Calculation


FTSE
Cumulative_P&L
0.3
0.2
0.1
0.0
−0.1
2013
2014
2015
2016
2017
2018
2019
2020
bindate

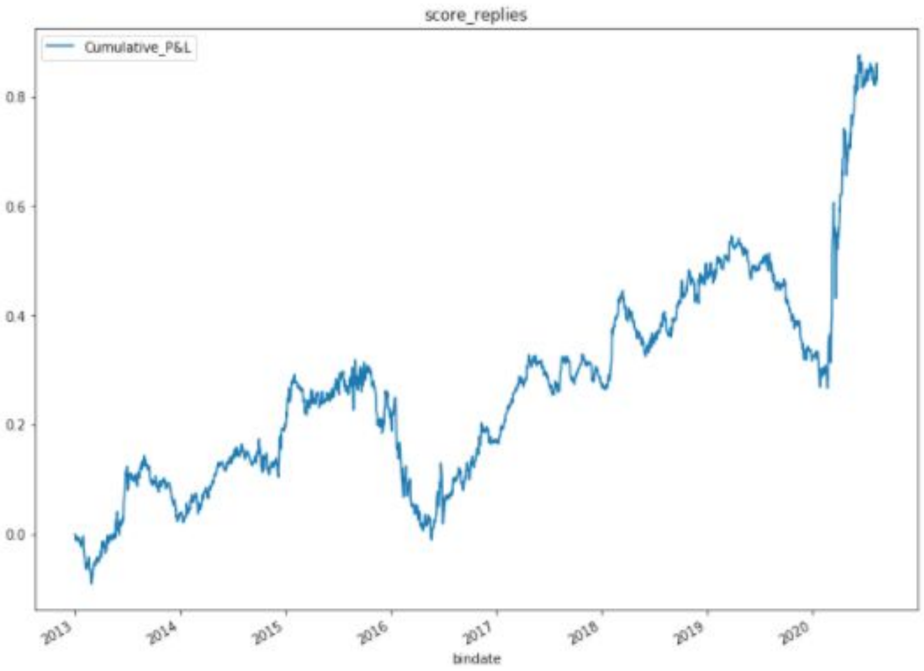

1,135.80 5 Oct 1984
8,000
6,000
4,000
2,000
0
1989
1994
1999
2004
2009
2014
2019

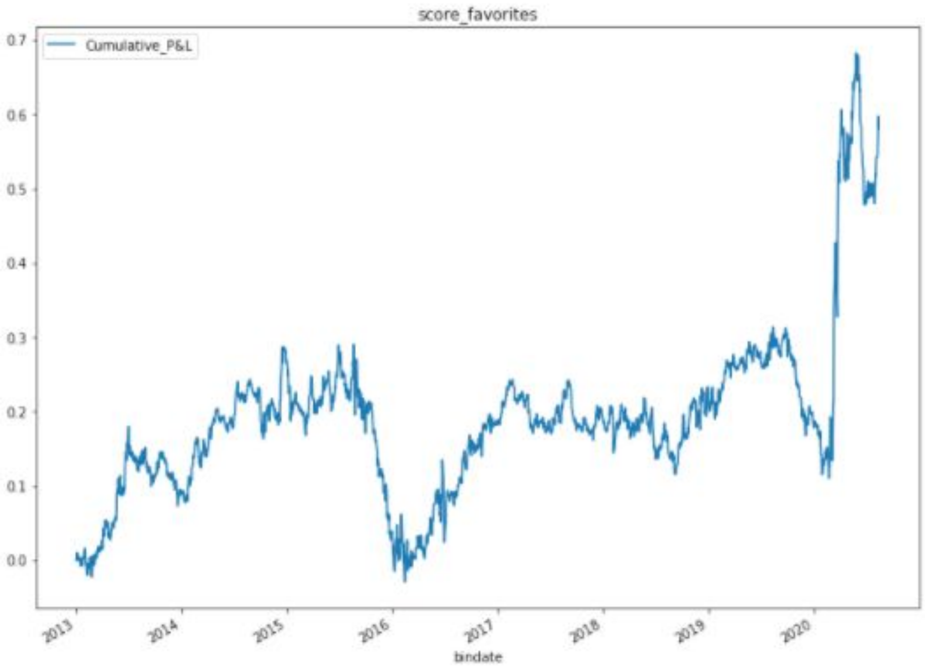
Graphs… Nice Visuals !





Continued... Score_Replies and Score_Favorites





Score_replies: 0.825

Score_favorites: 0.6

Score_Replies and Score_Retweets variables are Predictive and generate a positive P & L

This may be due to a Combination of Factors:

1. Tweets that get a lot of replies are likely to be opinions/speculations on future stock behavior.
2. The users that post make such speculations are also likely to have a lot of followers.
3. This can generate a discussion on twitter; the replies to the tweet could influence other users trading behaviour i.e. users who either read or replied to that tweet.

- It is highly probable that these opinions seep through and influence other users trading decisions.

- Correlation and Causation, could tweets influence trading patterns?(Retail investors)

(CAGR) Compound Annual Growth Rate

Get the annualized profit and calculate the Compound annual growth rate (CAGR)

Score Replies (CAGR): 8.24%

Score Favorites (CAGR): 6.38%

Bank of England: **0.1%** interest rate

Inflation Rate 2017: **2.56%**

$$\mathrm{CAGR} = \left(\frac{V_{\mathrm{final}}}{V_{\mathrm{begin}}}\right)^{1/t} - 1$$

CAGR = compound annual growth rate

V_{begin} = beginning value

V_{final} = final value

t = time in years

Sharpe Ratio

- Sharpe Ratio is used to help investors understand the return of an investment compared to its risk. The ratio is the average return earned in excess of the risk-free rate per unit of volatility or total risk. Volatility is a measure of the price fluctuations of an asset or portfolio.
- consistency of the return,

```
for var in ['score', 'score_retweets', 'score_replies', 'score_favorites', 'optimal_score']:
    pos = ftse13[var]
    print(var)
    pos = pos - pos.rolling(window=5,min_periods=1).mean()  #subtract bias
    pos = np.sign(pos)#convert to 1 or -1
    pnl = pos.shift(1) * ftse.Returns
    sharpe = np.sqrt(252) * pnl.mean()/pnl.std()#pnl standard profit and loss
    print(sharpe)
```

Formula and Calculation of Sharpe Ratio

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

where:

R_p = return of portfolio

R_f = risk-free rate

σ_p = standard deviation of the portfolio's excess return

score
0.1450019271904248
score_retweets
0.028753019956425315
score_replies
0.670240885675957
score_favorites
0.46919295607869493

Sharpe Ratio Grid Search/ Optimization

```python
wts = list(itertools.permutations([0.1, 0.2, 0.3, 0.4]))
wts = wts+[(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)]
lookbacks = [5] + list(range(22,22*13,22))

def calc_sharpe(lookback, wt):
    pos = wt[0]*ftse13.score + wt[1]*ftse13.score_retweets + wt[2]*ftse13.score_replies + wt[3]*ftse13.score_favorites
    pos = pos - pos.rolling(window=lookback, min_periods=1).mean()  #subtract bias, takes the average over a different
    # position vector long or short it
    # todays sentiment subtracted the last 3month
    pos = np.sign(pos)#convert to 1 or -1
    pnl = pos.shift(1) * ftse.Returns
    sharpe = np.sqrt(252) * pnl.mean()/pnl.std()#pnl standard profit and loss
    return sharpe

maxsharpe = 0
maxvars = None
for wt in wts:
    for lookback in lookbacks:
        sharpe = calc_sharpe(lookback, wt)
        if sharpe > maxsharpe:
            maxsharpe =sharpe
            maxvars = (wt, lookback)

print(maxsharpe)
print(maxvars)
```

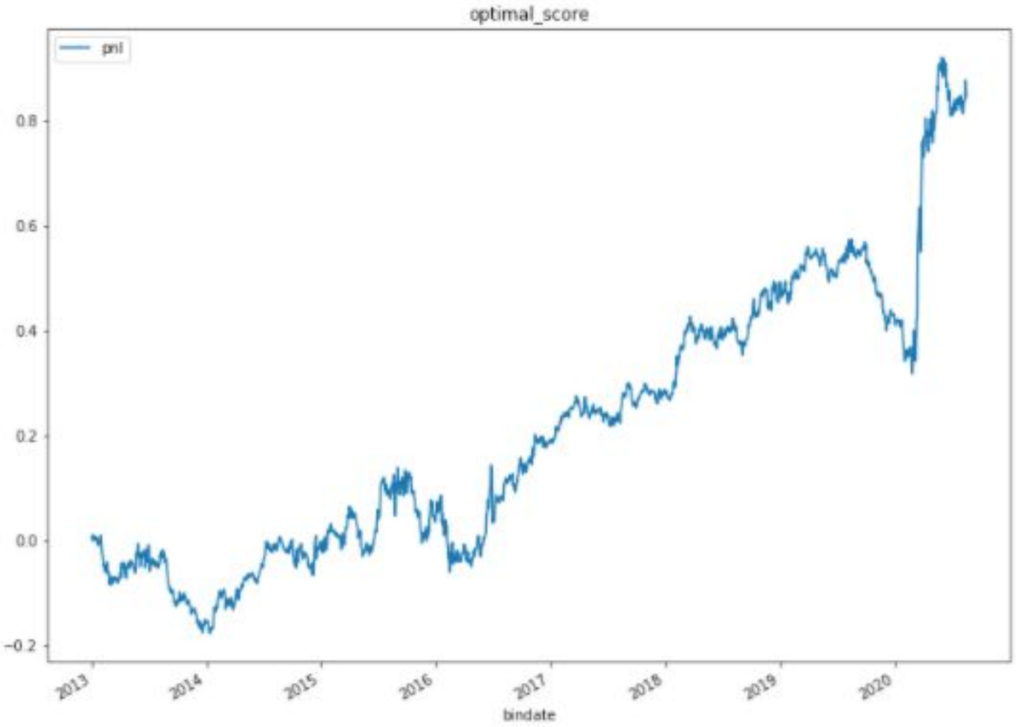
Optimal Score: **0.6836777690933067**
Optimal weights for variables: (**0.3***score, **0.2***score_retweets, **0.4***score_replies, **0.1***score_favorites)
Optimal Lookback time period = **5** working days/ week

Adding a new Column... Optimal_Score

```
ftse13['optimal_score'] = 0.3*ftse13.score + 0.2*ftse13.score_retweets + 0.4*ftse13.score_replies + 0.1*ftse13.score_fa
```

bindate	score	score_retweets	score_replies	score_favorites	retweets	favorites	replies	Close	Returns	optimal_score
2013-01-01	0.050727	-0.208900	0.320875	0.000000	18.0	0.0	4.0	5897.81	0.000000	0.101788
2013-01-02	0.070267	0.174218	0.293008	0.095222	99.0	9.0	26.0	6027.37	0.021967	0.182649
2013-01-03	0.211909	0.238970	0.120747	0.303846	94.0	13.0	34.0	6047.34	0.003313	0.190050
2013-01-04	0.050287	0.122616	0.178419	0.478550	57.0	4.0	26.0	6089.84	0.007028	0.158832
2013-01-07	0.178135	0.054998	0.154537	0.137562	130.0	26.0	19.0	6064.58	-0.004148	0.140011
...	...	...	...	...	...	...	...	...	...	...
2020-08-10	0.146385	0.255829	0.276632	0.279631	242.0	693.0	97.0	6050.59	0.003052	0.233697
2020-08-11	0.179729	0.426062	0.210567	0.382079	128.0	490.0	91.0	6154.34	0.017147	0.261566
2020-08-12	0.116080	0.210003	0.188777	0.246521	298.0	936.0	129.0	6280.12	0.020438	0.176987
2020-08-13	0.133928	0.181326	0.319378	0.307921	359.0	1347.0	156.0	6185.62	-0.015047	0.234987
2020-08-14	0.236990	0.612014	0.606838	0.611605	505388.0	793472.0	7739.0	6090.04	-0.015452	0.497395

Optimal_Score



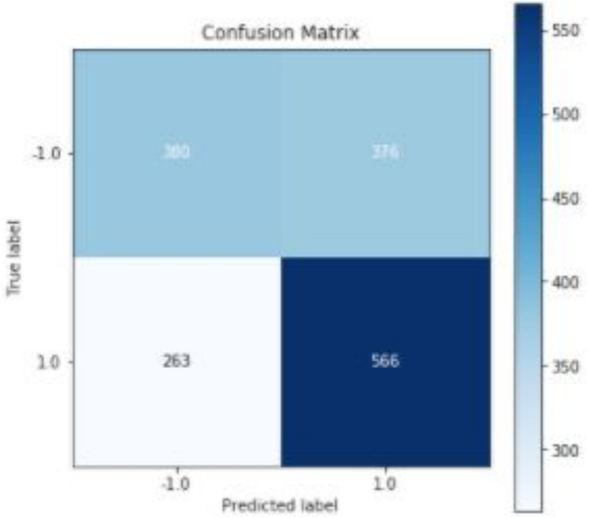
Optimal_Score (CAGR): **8.43%**

Classification

Target Variable: Position taken(long or short)/(1 or -1)

Logistic Regression

Best Score: 0.5981072555205047



Findings

1. Twitter sentiment reflects the mood of the market. **yes**

2. Twitter can be used to predict future stock market movements. **yes**

The results indicate that Twitter data is a suitable source to help understand and forecast future stock market movements

Sentiment extracted from Twitter has significant predictive power for predicting the direction for the Returns.

Conclusions

The Trading Strategy works, it is possible to generate large profits over a large period of time.

Next Steps

- Apply the Trading Strategy to other indices, Gold(inverse correlation)
- Tune the model
- Sell the Strategy???$$$
- There is already some documentation on leveraging Twitter Data to help predict stock market prices.
- Sentiment data is also collected from other sources, and is used as a tool for investors.
- 75% of trading is Automated

Thank you Dan and Christoph!!