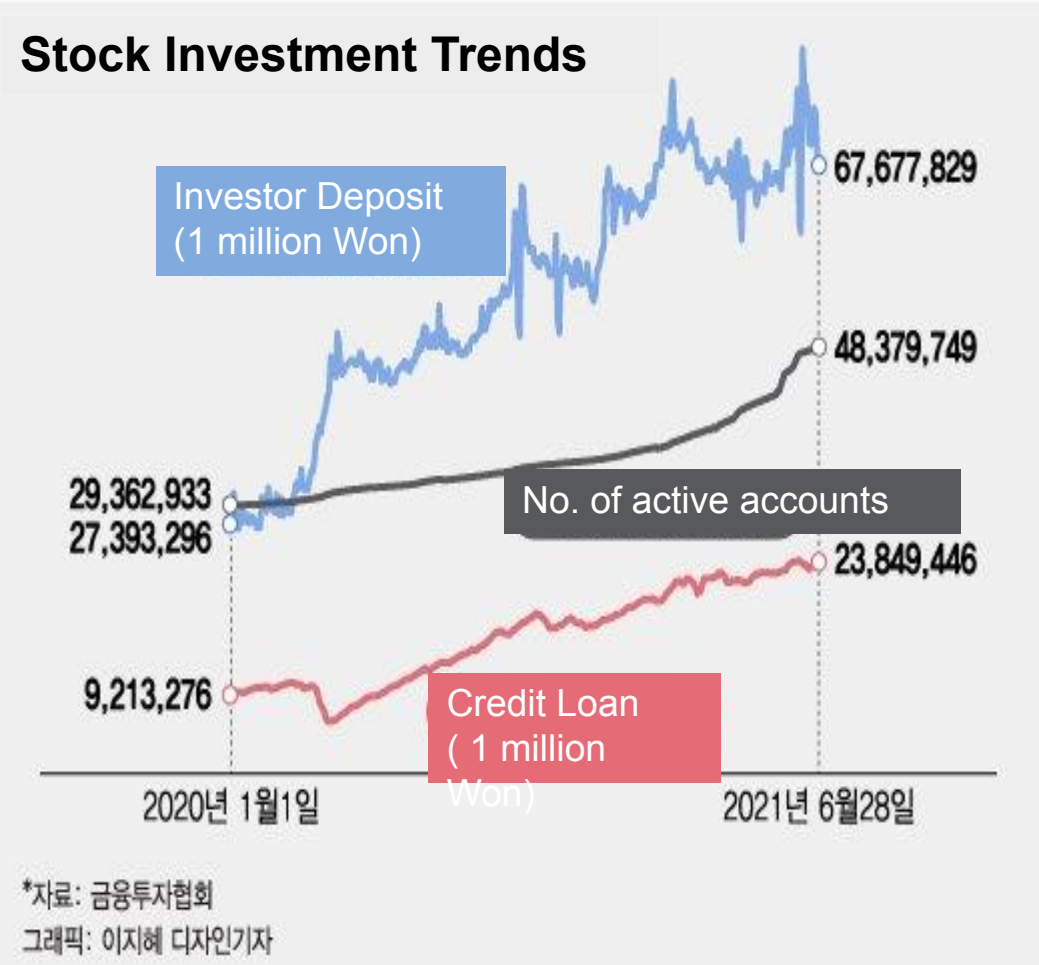
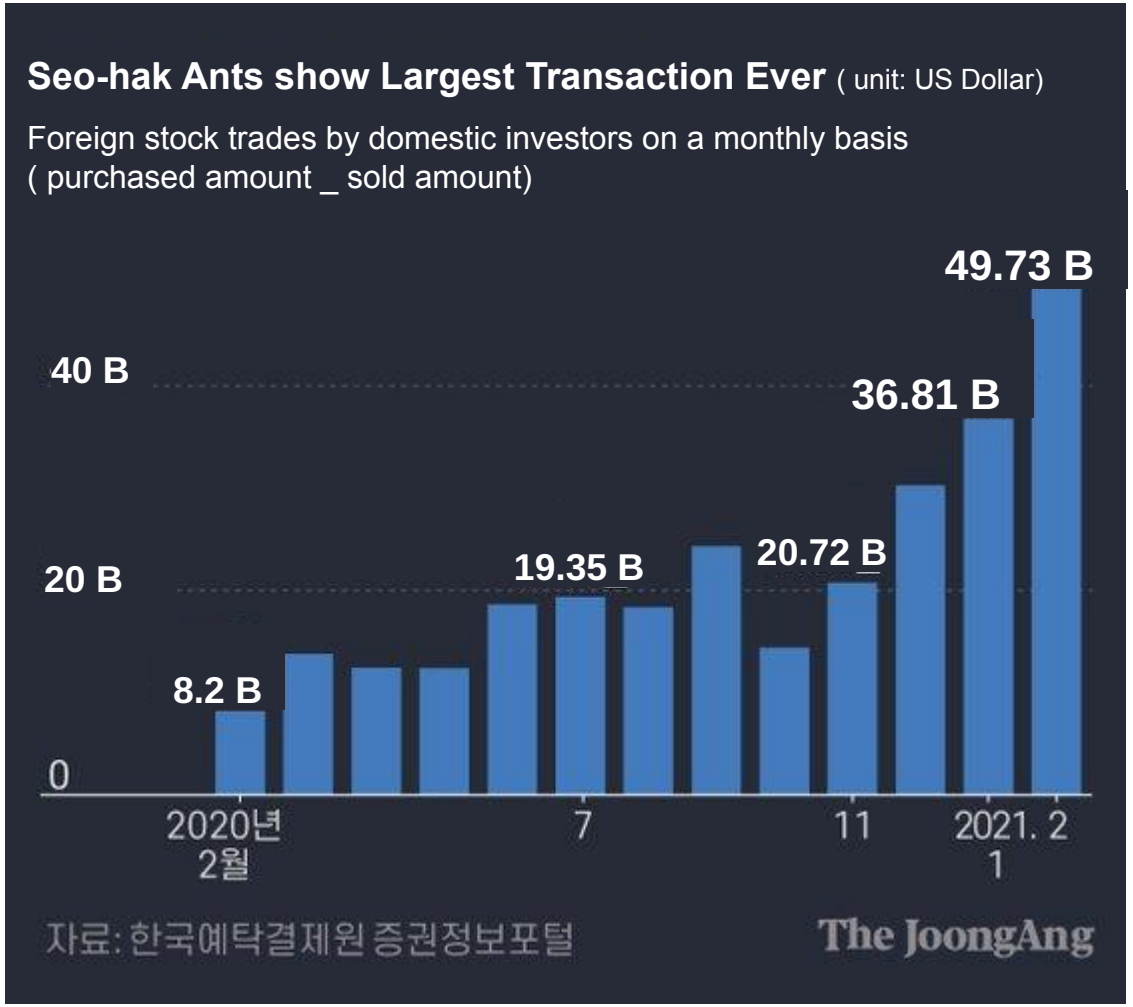
DATUM

Group2: Stock_Analysis

Contents

Introduction





Research Methodology - Twitter sentiment analysis

*** Polarity Detection**
Tried to detect the emotion of each tweets by using polarity detection. The tweets were divided into three groups based on the sentiment of the pre-processed tweets.
negative words were given '-1' as scores, positive ones '+1', and '0' for netural words.

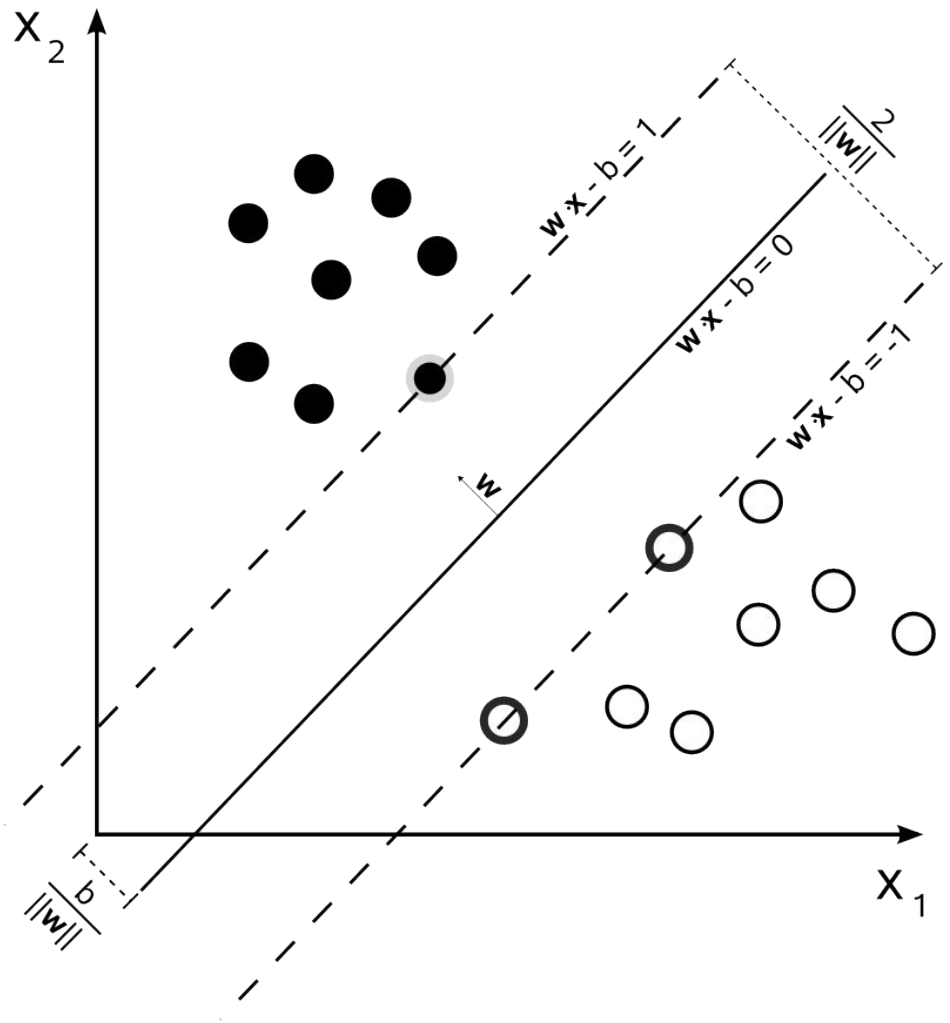
Stock Price Prediction Model

- Scraped Twitter to analyze sentiment changes on a company/stock

Module used: SNScrapy, NLTK

- scraped more than **50,000 tweets** per stocks (ticker used: TSM, RCL)
- Tweets containing company name or ticker as **hashtags** were scraped
  i.e. for Apple Inc. #Apple OR #AAPL
- **Historical Data Length**: 3 years
- Used ***NLP*** for preprocessing & applied ***Polarity Detection*** to analyze the sentiment on a ticker(stock).

Research Methodology – Support Vector Machine



3 Class

Positive(+1), Neutral(0), Negative(-1)

Assumption

Neutral tokens will be located near the division line.

Process

1. Split the data into training set / test set using sklearn
2. Create SVM classifier by using SVC from sklearn
3. Used the score to check the performance of the classifier

Data Overview – raw tweets data

tweets_df

	Datetime	Tweet id	Text	Username
0	2021-11-09 23:57:56+00:00	1458222354685997062	#RichardFain steps down as CEO of #RoyalCaribb...	scottlara1961
1	2021-11-09 23:50:52+00:00	1458220577659916291	@RoyalCaribbean Your customer service departme...	DavidFr98098022
2	2021-11-09 23:43:06+00:00	1458218621570715650	@danni_is_woke Hey, Danni. I'm sorry to hear. ...	RoyalCaribbean
3	2021-11-09 23:41:38+00:00	1458218253696733185	@StephAndSelves @RoyalCaribbean Our problem is...	DavidFr98098022
4	2021-11-09 23:40:51+00:00	1458218054693752839	Have you checked out our giveaways happening a...	LuckysLoungeLV
...	...	...	...	...
99996	2020-03-09 22:49:09+00:00	1237148445955129345	@Examinwithme @RoyalCaribbean @CruiseNorwegian...	Naturally_Kelz
99997	2020-03-09 22:48:12+00:00	1237148207961870338	@bwk1992 @RoyalCaribbean Exactly. Rescheduling...	Naturally_Kelz
99998	2020-03-09 22:44:58+00:00	1237147392643825672	@RoyalCaribbean We purchased the travel insura...	AngelicaLakhani
99999	2020-03-09 22:41:23+00:00	1237146491078279168	@CrucerosPR @RoyalCaribbean ...dices que el Ca...	daya_ojitos
100000	2020-03-09 22:38:01+00:00	1237145646462795778	@RoyalCaribbean thinking of cancelling our Apr...	Joe4alb

Data set Overview

The Japanese government will establish a legal framework for subsidizing new domestic plants for advanced semiconductors, starting with TSMC's planned facility in Kumamoto.
#TSMC #Taiwan
https://t.co/3rtOitA6vH

@SIMONLUI11 @SmartTaipei @Fannyi5 @Reginalplau You clearly don't understand semiconductor industry.

1. Most of design is supplied by #ARM.
2. Chip itself is fabbed at #TSMC.
3. When the US sanctions #Alibaba, that's the end of that chip, just like #Huawei.

https://t.co/hFu2RAT73Z

Examples of Tweets Scraped

Module : snscrapy (source: https://www.kaggle.com/antonhansson/fetch-tweets-covid-19-vaccine)

- 100,000 tweets, 640 days length
- Tweets about Royal Caribbean Cruise, TSMC (Ticker: RCL, TSM)
- Used hashtags to search related tweets i.e. for RCL we used #RoyalCaribbean OR #RCL

Data Overview – preprocessed data

```
RCL_polarity_real.head()
```

	ticker	date	tweets	neg	neu	pos	compound
0	RCL	2021-11-09	richardfain step ceo royalcaribbean httpstcofp...	0.000	1.000	0.000	0.0000
1	RCL	2021-11-09	royalcaribbean customer service department ver...	0.333	0.573	0.093	-0.6908
2	RCL	2021-11-09	danni_is_woke hey danni I m sorry hear reach a...	0.138	0.745	0.117	-0.0516
3	RCL	2021-11-09	stephandselve royalcaribbean problem stateroom...	0.309	0.631	0.060	-0.7579
4	RCL	2021-11-09	check giveaway happen spark location grand pri...	0.000	0.423	0.577	0.9738

Module : NLTK, re

- removed non-english languages during preprocessing
- User ID and Username was used to detect and remove duplicated data

Data Overview – Results

	usercreatedts	text
0	2009-07-27 06:41:57	Taiwan #Semiconductor Manufacturing Co. #TSMC ...
1	2015-09-25 18:44:17	TSMC says it will build first Japan chip plant...
2	2019-08-31 02:16:48	Preparing the EU #ChipsAct, important meeting ...
3	2020-09-10 14:06:32	TSMC says it will build first Japan chip plant...
4	2021-04-20 22:40:59	TSMC says it will build first Japan chip plant...
...	...	...
30754	2021-04-23 22:24:01	@SahilBloom @horwitzjosh @benthompson @ShaneAP...
30755	2019-12-28 00:16:49	@WEIWEIDAI4 Wait until when they get pressured...
30756	2020-02-27 14:12:39	TSMC founder:\n\n"Intel CEO Pat Gelsinger 'a v...
30757	2011-08-08 07:24:32	@iingwen Please answer them by introducing chi...
30758	2012-10-15 05:13:37	#TSMC may get a 50b$ deal with #india #semic...

Example dataset: from 2007-03-12 to 2021-11-10 randomly chose 247 days (ticker:TSMC)

original tweet retweeted multiple times

	usercreatedts	text	senti_score
19001	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
12018	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
12427	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
28087	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
12836	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
13246	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
5508	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
27674	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
13656	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857
1033	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	0.342857

Sorted by Created time(including retweets)

```
tweets=data[['usercreatedts','text']]
tweets=tweets.sort_values('usercreatedts')  #did not drop the duplicates

blobs=[]
for tweet in tweets['text']:
  blob=TextBlob(tweet)
  blobs.append(blob)

sentiment_score=[]
for blob in blobs:
  # print(f"- sentiment score {blob.sentiment.polarity}: {blob}")
  sentiment_score.append(blob.sentiment.polarity)
tweets['senti_score']=sentiment_score
```

Multiple duplicated due to 'retweets'

→ did not remove the duplicates because the number of retweets shows how many others agree or show interest with that opinion.

Module: TextBlob, VaderSentiment

- Originally used TextBlob
- Improved the sentiment analysis by using VADER sentiment!

Data Overview – Results

```
pddate=pd.to_datetime(tweets['usercreatedts'])
tweets['usercreatedts']=pddate.dt.date  #remove the time
data=tweets.groupby('usercreatedts').sum()
```

usercreatedts	senti_score
2007-03-12	25.714286
2008-02-15	0.390000
2008-04-11	0.000000
2008-05-07	-9.375000
2008-05-23	30.000000
...	...
2021-10-24	-9.375000
2021-10-25	-63.375000
2021-11-01	26.250000
2021-11-06	25.714286
2021-11-08	11.996753

	senti_score
count	247.000000
mean	12.207688
std	32.725793
min	-63.375000
25%	0.000000
50%	5.113636
75%	25.714286
max	378.523728

sentiment analysis is easily done for each tweets

Remaining question: how do we calculate the daily sentiment?

i.e. There could be a mix of negative and positive sentiments in a day

- Currently using simple summation
  (n: number of tweets generated that day
  (i: index number)

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

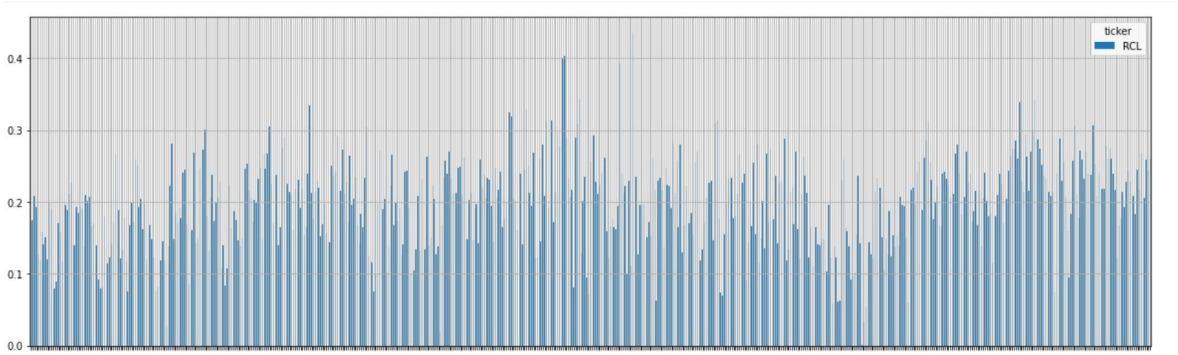
- Discussing how to apply appropriate weight

$$W = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

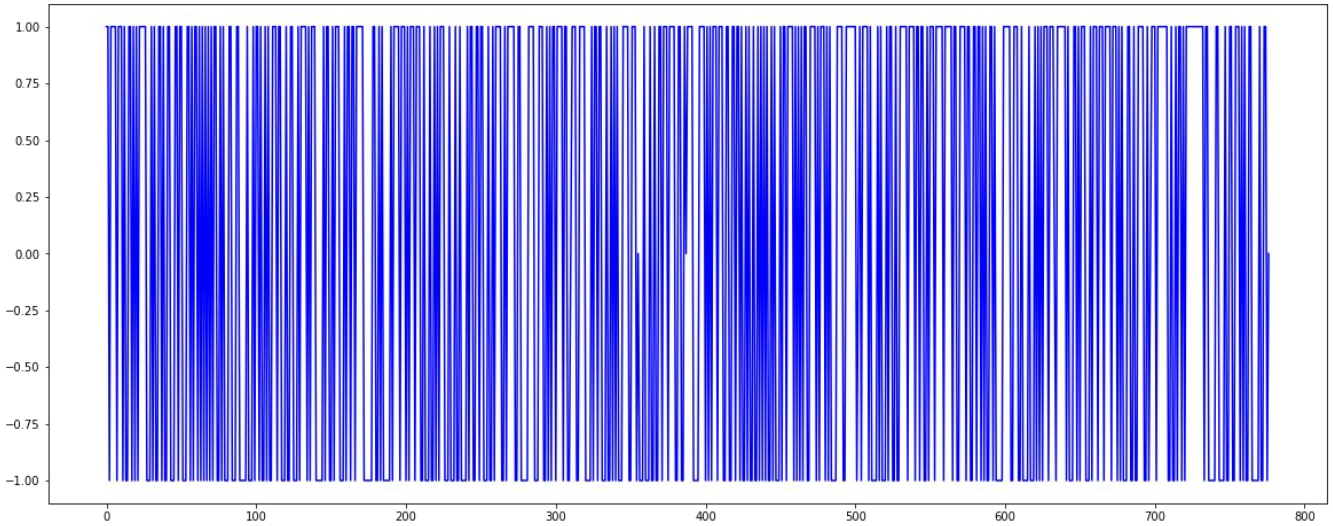
  Considered using the amount of hearts a tweet recieved, but it would be hard to determine when the heart was sent, It could have been 'liked' two years after!

  → Suggestions are welcomed!

Data Overview – Results



Sentiment score of RCL

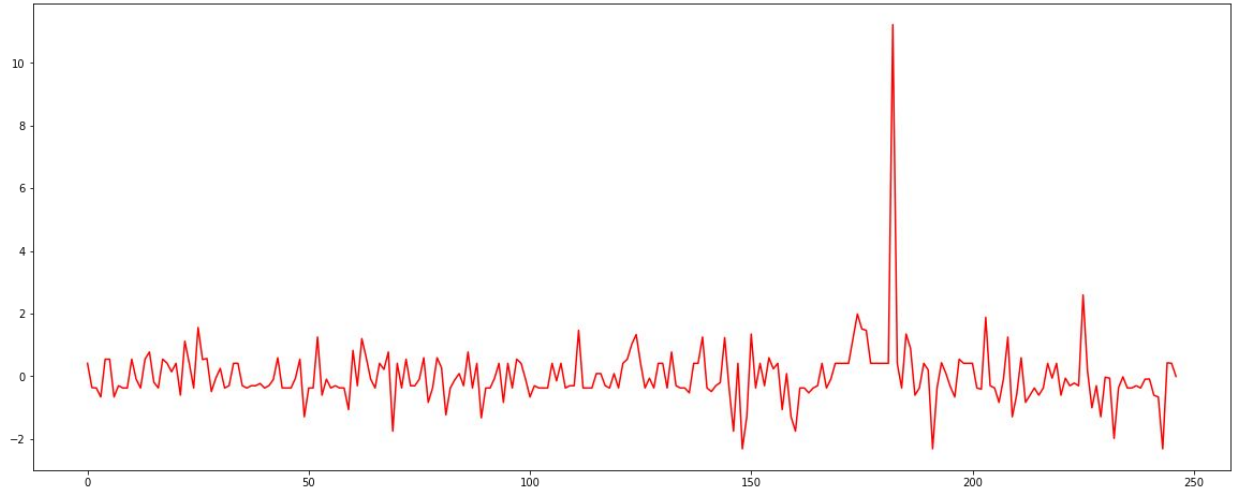


Above: Price change of TSMC
Below: Standardized Daily Sentiment score of TSMC

	usercreatedts	text	blob	senti_score
19001	2007-03-12 22:34:58	Preparing the EU #ChipsAct, important meeting ...	(P, r, e, p, a, r, i, n, g, , t, h, e, , E, ...	0.342857

Trying to find the correlation between rate of change & sentiment

No significance found yet.



Limitations

1. There is a distinct disparity of publicly available tweets between well known company stocks and lesser-known stocks.

   i.e. more than 50,000 tweets about Facebook, Microsoft are generated per day

2. The dictionary in NLTK lacks information on stock related acronyms and internet slangs. i.e. P.E(Price to Earnings), stonks(refers to stocks that cost financial loss)

   → User dictionary must be made for supplementation

3. Sentiment of the company is not the only factor that affects stock price. Must consider technical factors, and company fundamentals as well.

4. Weighting the sentiments.

Limitations

1

There is a distinct disparity of publicly available tweets between well known company stocks and lesser-known stocks.

→ **Solution #1. Limit the research scope to stocks listed on S&P 500 or Dow Jones**
→ **Solution #2. Supplement lesser-known stocks by additionally scraping news articles**

2

The dictionary in NLTK lacks information on stock related acronyms and internet slangs. → User dictionary must be made for supplementation

→ **Complement user dict. using Investopedia and other online sites for financial terms.**

3

Sentiment of the company is not the only factor that affects stock price. Must consider technical factors, and company fundamentals as well.

→ **Improve model accuracy by adding more indicators such as PER, MACD etc.**

4

Weighting the sentiments.

Thank you