Sentiment Analysis on Forums to Predict Stock Market

Hong-Wen, Yung-Yu

October 2022

I. Data Explore

Data Type: 923,673 Tweets data
Time period: 2020/04/09 - 2020/07/16 (only got 77 out of 99 days)

Preprocessing Step:

1. Convert to lower case
2. Convert www.* or https?://* to URL
3. Convert @username to USER
4. Remove additional white spaces
5. Replace word with word
6. Trim front and back
7. Remove stop words

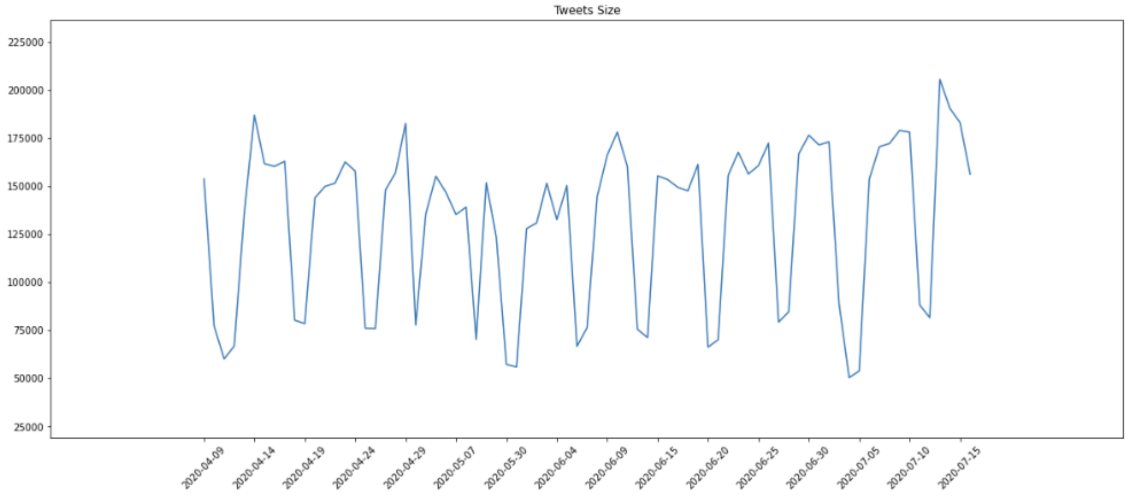


Figure 1: Tweets size chart

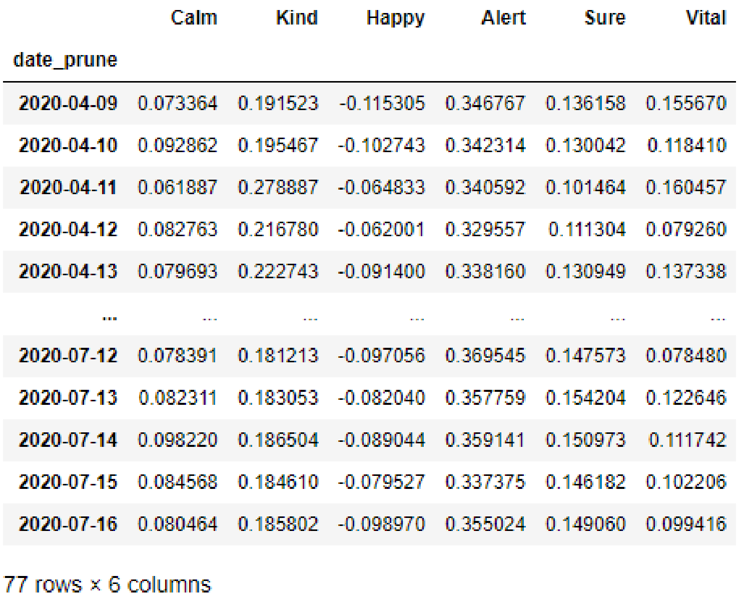
II. Sentiment analysis

1. GPOMS

Sentiment Types:

1. Calm: composed/anxious
2. Kind: agreeable/hostile
3. Happy: elated/depressed
4. Alert: confident/unsure
5. Sure: clearheaded/confuse
6. Vital: energetic/tired

Analysis result: ”Alert” is the highest estimator among 77 days



	Calm	Kind	Happy	Alert	Sure	Vital
date_prune						
2020-04-09	0.073364	0.191523	-0.115305	0.346767	0.136158	0.155670
2020-04-10	0.092862	0.195467	-0.102743	0.342314	0.130042	0.118410
2020-04-11	0.061887	0.278887	-0.064833	0.340592	0.101464	0.160457
2020-04-12	0.082763	0.216780	-0.062001	0.329557	0.111304	0.079260
2020-04-13	0.079693	0.222743	-0.091400	0.338160	0.130949	0.137338
...	...	...	...	...	...	...
2020-07-12	0.078391	0.181213	-0.097056	0.369545	0.147573	0.078480
2020-07-13	0.082311	0.183053	-0.082040	0.357759	0.154204	0.122646
2020-07-14	0.098220	0.186504	-0.089044	0.359141	0.150973	0.111742
2020-07-15	0.084568	0.184610	-0.079527	0.337375	0.146182	0.102206
2020-07-16	0.080464	0.185802	-0.098970	0.355024	0.149060	0.099416

77 rows × 6 columns

Figure 2: GPOMS result

2. Opinion Finder

4 types of lexicons:

1. Weak negative: 1175 words, 1 point
2. Strong negative: 3737 words, 1.5 point
3. Weak positive: 1129 words, 1 point
4. Strong positive: 2180 words, 1.5 point

Processing Step:

1. Calculate each day’s score
2. Positive point - Negative point
3. Minmax transform
4. Apply funciton: $y = 2x - 1$

Scale after transform: -1 - 1

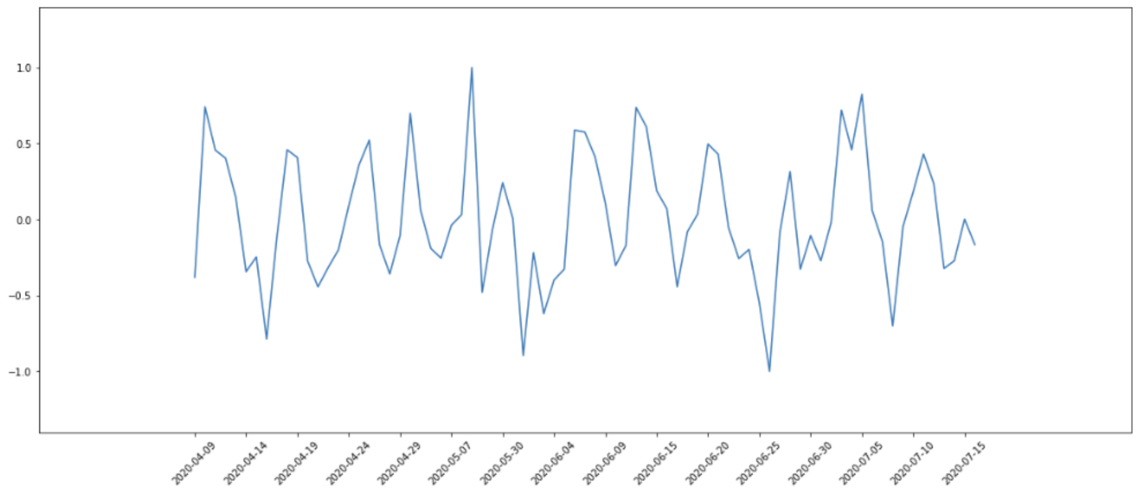
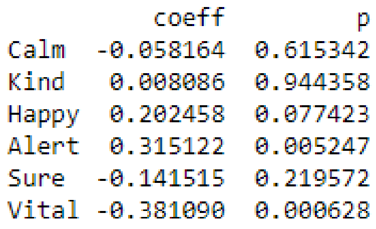


Figure 3: Opinion Finder result

III. Granger Analysis



```
            coeff         p
Calm  -0.058164  0.615342
Kind   0.008086  0.944358
Happy  0.202458  0.077423
Alert  0.315122  0.005247
Sure  -0.141515  0.219572
Vital -0.381090  0.000628
```

Figure 4: Regression between Opinion Finder and GPOMS

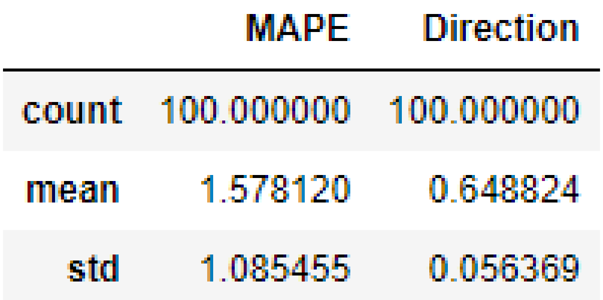
Granger Causality test is used to determine whether or not one time series is useful for forecasting another.

Sentiment with causality: Sure(p-value=0.05), Alert(p-value=0.11)

Highest granger gap days(align with best modal):

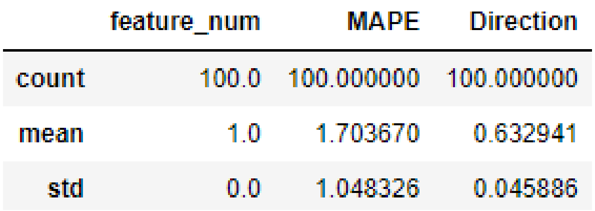
- Opinion Finder: 5 days (0.11)
- Alert: 5 (0.11)
- Sure: 1 (0.05)
- Alert + Sure: 4 (0.08)
- Alert + Vital: 4(0.11)
- Alert + Sure + Vital: 4(0.08)
- Happy + Alert + Sure + Vital: 3(0.14)

IV.Forecast Result



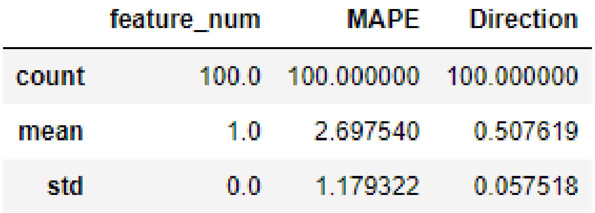
	MAPE	Direction
count	100.000000	100.000000
mean	1.578120	0.648824
std	1.085455	0.056369

Figure 5: Opinion Finder



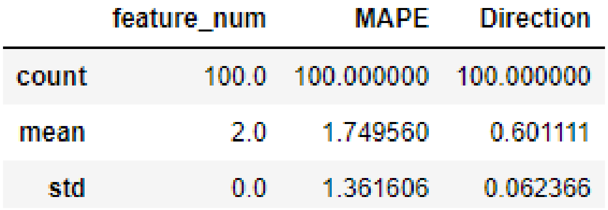
	feature_num	MAPE	Direction
count	100.0	100.000000	100.000000
mean	1.0	1.703670	0.632941
std	0.0	1.048326	0.045886

Figure 6: Alert



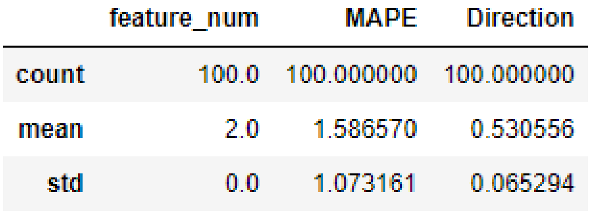
	feature_num	MAPE	Direction
count	100.0	100.000000	100.000000
mean	1.0	2.697540	0.507619
std	0.0	1.179322	0.057518

Figure 7: Sure



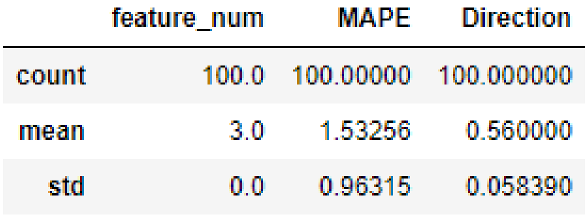
	feature_num	MAPE	Direction
count	100.0	100.000000	100.000000
mean	2.0	1.749560	0.601111
std	0.0	1.361606	0.062366

Figure 8: Alert + Sure



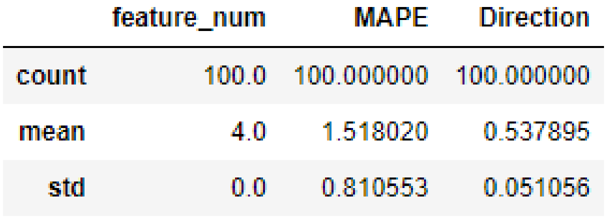
	feature_num	MAPE	Direction
count	100.0	100.000000	100.000000
mean	2.0	1.586570	0.530556
std	0.0	1.073161	0.065294

Figure 9: Alert + Vital



	feature_num	MAPE	Direction
count	100.0	100.00000	100.000000
mean	3.0	1.53256	0.560000
std	0.0	0.96315	0.058390

Figure 10: Sure + Alert + Vital



	feature_num	MAPE	Direction
count	100.0	100.000000	100.000000
mean	4.0	1.518020	0.537895
std	0.0	0.810553	0.051056

Figure 11: Sure + Alert + Vital + Happy