MGT 6090
Management of Financial Institutions

Assignment 12
Event Studies and Sentiment Analysis using Python
Week 12

Parth Shah
GT ID: 903619858
MS QCF
pshah399@gatech.edu

Python Assignment 12: Event Studies and Sentiment Analysis using Python

1 Data Downloading and preparation

1. Download Data from the SEC We download 50 random 8-Ks for each year-quarter from the SEC website for the time period 1995:Q1 through 2020:Q4.

We clean the master.idx file (after unzipping), in order to remove the first few lines use regex to filter the forms so that you can keep only the 8-K filings Extract 50 random lines or companies for each year-quarter Extract the path name or link for the 8-K download Download the corresponding 8-Ks Create file that keeps track of the company identifier (CIK) and the 8-K filing date

2 Compute abnormal stock returns and abnormal trading volume

Around 8-K filings. Using the file that contains information on CIKs and 8-K filing dates

the windows are, in relation to the event date, 0, -1,+1, -2,+2, -3,+3, -5,+5. Then, we define the daily abnormal stock return of firm i on day t, , as the residual estimated from the market model: $AR_i = R_i - (\alpha + \beta * R_m)$

Then, we calculate the CAR, which is the 3-day CAR in the 3-day window centered around the day of the 8-K filing. CAR0 - CAR5 Once these CAR's are computed, we compute descriptive statistics of the CAR along with the plotting the distribution.

3 Observations from Event Studies

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	1628.000000	1628.000000	1628.000000	1628.000000	1628.000000	1628.000000
mean	0.453408	-0.001725	-0.002960	-0.001682	-0.002135	0.000221
std	1.190342	0.053029	0.090100	0.124118	0.152190	0.205021
min	-6.216906	-0.608473	-0.951079	-0.579308	-0.742702	-1.185558
25%	-0.303583	-0.020571	-0.043136	-0.059559	-0.077346	-0.101873
50%	0.404547	-0.001185	-0.000330	-0.001358	-0.000699	-0.002357
75%	1.143917	0.018676	0.040080	0.060623	0.072156	0.098368
max	5.961652	0.282731	0.405881	1.598387	1.648797	1.647928

Figure 1: Descriptive Stats for Events studies

As shown in the descriptive stats, ATO has the highest range, standard deviation, and mean. The range of distribution for CAR0-CAR5 the distribution is from around -1 to 1.5.

CAR0 has the lowest max, with max increasing as we increase the day ranges from 0 to 5. and vice versa with the min of CAR.

The mean of CAR0 is low at around 0 and then CAR1, CAR2, CAR3 are just above 0 and CAR5 is below 0. The mean stays around 0.003, which goes in line with how 8-K stock announcements

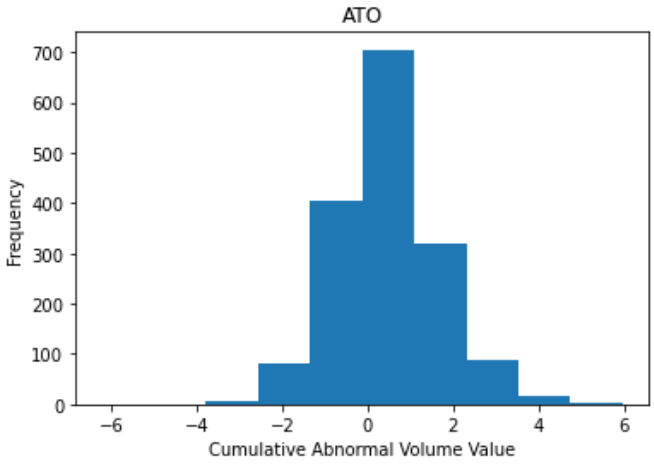

Figure 2: Turnover histogram plot

are. This tells us the the information of the filling is absorbed within 1-3 days of filling the 8k centered around the filing date.

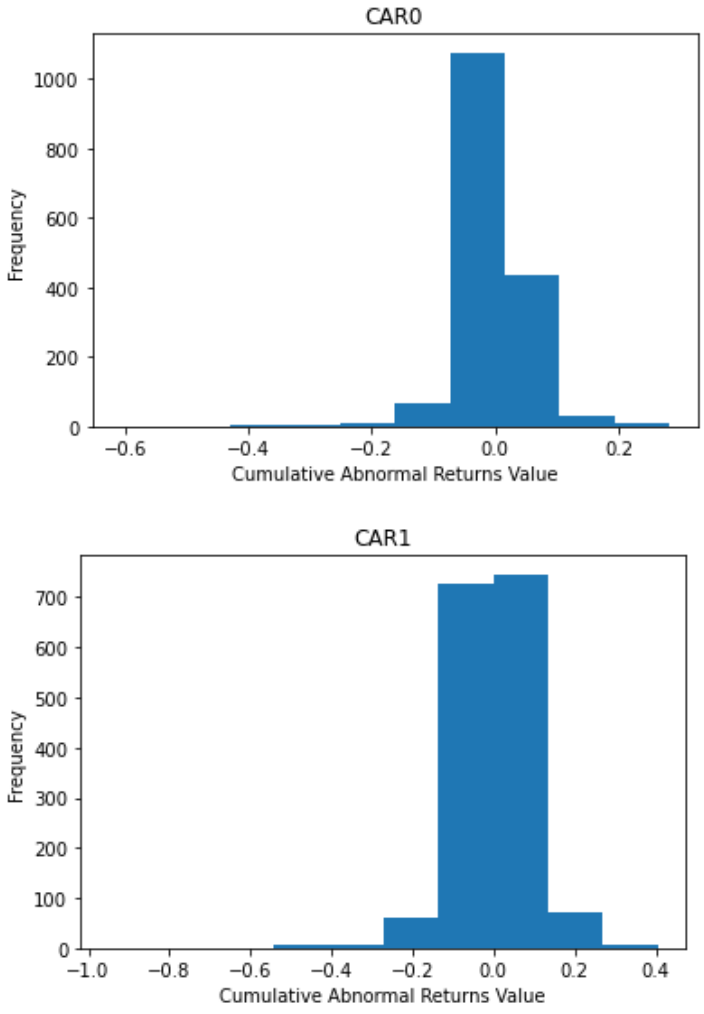

Figure 3: Cumulative Abnormal Returns

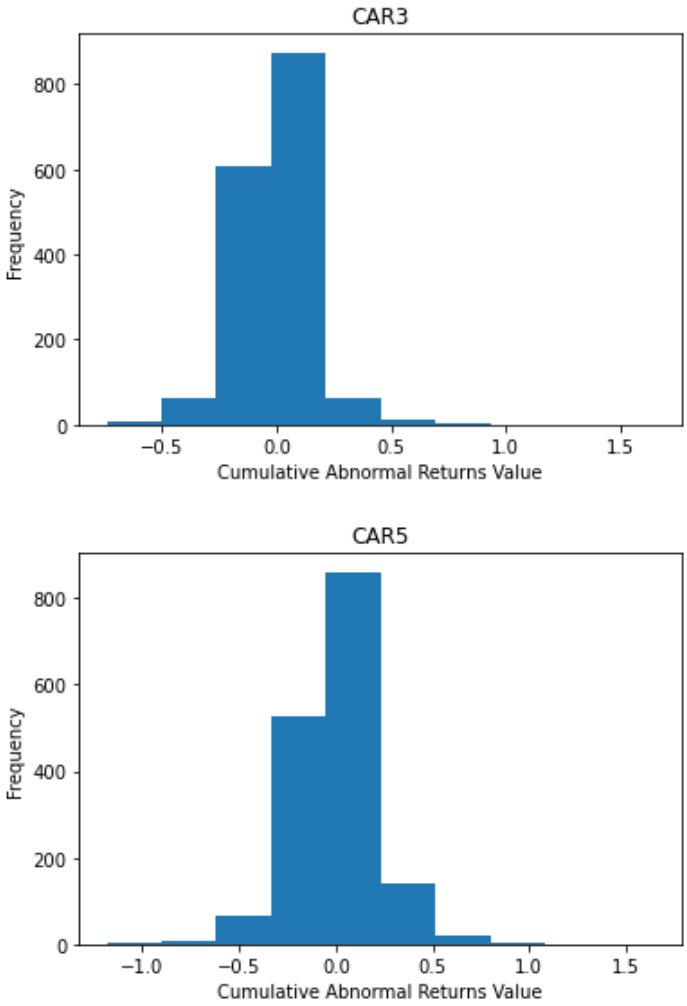

Figure 4: Cumulative Abnormal Returns

4 Rudimentary Sentiment Analysis

We use the words list from Bill McDonald's website: For each downloaded 8-K filing we calculate the difference bewteen the number of positive words and the number of negative words, and scale this difference by the total number of words in the document. Sort the measure constructed in the above step into quintiles at an annual frequency so that the lowest quintile represents the most negative document.

5 Less Rudimentary Sentiment Analysis

In the previous step, the tone of the 8-K filing was captured through simple counts of the number of negative and positive words in the filing. However, this sentiment measure is fairly imprecise. Now we use natural language processing (NLP) in order to more accurately capture the sentiment of the 8-K filing.

For each downloaded 8-K filing we identify the informational component of the filing by ignoring header section.

We then break the paragraph of information into sentences, and tokenize these individual sentences and assign a tone value to each sentence.

We use the VADER sentiment analysis toolkit from NLTK as a starting point. NLTK assigns weights of negativity, neutrality, and positivity to input data. The weights sum up to 1. We shall use the compounded weight assigned to each sentence as a measure of the sentence's tone and calculate the average tone value of all sentences in the document. We now sort this measure into quintiles at an annual frequency.

6 Comparison betweeen Rudimentary and advanced Sentimental Analysis

We show the descriptive stats for each quintile for both rudimentary and advanced methods to compare and ocntrast the two methods results.

The histogram plots can be seen in the html file and ipynb file.

Quintile 0 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	484.000000	484.000000	484.000000	484.000000	484.000000	484.000000
mean	0.344270	0.001488	0.002227	0.001082	0.009045	0.014483
std	1.195218	0.058430	0.088875	0.113027	0.140683	0.178200
min	-3.706080	-0.265067	-0.293253	-0.529028	-0.538037	-0.890956
25%	-0.430218	-0.017639	-0.034690	-0.045453	-0.052425	-0.070795
50%	0.256988	0.000995	-0.000078	-0.001849	0.004508	0.008777
75%	0.943709	0.017754	0.032122	0.041867	0.071255	0.089881
max	6.277401	0.524418	0.589578	0.643714	0.886869	0.992884

Figure 5: Rudimentary quintile 0 Descriptive Stats

Lowest Quintile i.e. quintile 0, we can see that less rudimentary descriptive stats are far more significant then rudimentary step. So very negative documents are more easily identified by the nlp model and mean CAR around those fillings is higher than the ruidmentary method.

Quintile 0 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	483.000000	483.000000	483.000000	483.000000	483.000000	483.000000
mean	0.528637	0.005496	0.007820	0.014614	0.016472	0.022459
std	1.206937	0.055313	0.084479	0.096535	0.116524	0.151900
min	-3.242131	-0.235830	-0.419506	-0.283011	-0.361848	-0.421659
25%	-0.236053	-0.015323	-0.025220	-0.036808	-0.045794	-0.070403
50%	0.427117	0.002203	0.002125	0.005717	0.007408	0.008928
75%	1.088289	0.020359	0.034072	0.049842	0.064604	0.087731
max	6.277401	0.446100	0.613460	0.662016	0.688933	0.809367

Figure 6: Less Rudimentary quintile 0 Descriptive Stats

Quintile 1 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000
mean	0.455878	0.002387	0.004039	0.013553	0.016888	0.019192
std	1.194844	0.053733	0.102804	0.124602	0.144285	0.174864
min	-3.706080	-0.235830	-0.575450	-0.578227	-0.582609	-0.512568
25%	-0.322593	-0.017991	-0.033296	-0.037856	-0.047962	-0.071671
50%	0.338747	0.001587	0.001844	0.006473	0.009348	0.006401
75%	1.097727	0.019214	0.033981	0.047897	0.062508	0.093652
max	4.218503	0.446100	0.980715	1.058998	1.080871	1.080871

Figure 7: Rudimentary quintile 1 Descriptive Stats

Quintile 1 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	483.000000	483.000000	483.000000	483.000000	483.000000	483.000000
mean	0.530243	0.001478	0.007918	0.013434	0.021244	0.022922
std	1.168872	0.058945	0.106706	0.138768	0.156024	0.191401
min	-3.062761	-0.486894	-0.612378	-0.860699	-0.942553	-1.056260
25%	-0.264454	-0.020037	-0.031471	-0.040549	-0.046511	-0.073189
50%	0.371430	0.001166	0.003331	0.005903	0.007610	0.011897
75%	1.170061	0.018722	0.036757	0.053996	0.073957	0.110212
max	6.277401	0.363340	0.980715	1.058998	1.080871	1.080871

Figure 8: Less Rudimentary quintile 1 Descriptive Stats

For Quintile 1, the difference is not that stark as we move from tighter CAR period to less tighter. But we see that less rudimentary way is slighlty more better.

Quintile 2 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	638.000000	638.000000	638.000000	638.000000	638.000000	638.000000
mean	0.484084	0.002700	0.005952	0.010930	0.011539	0.018126
std	1.178666	0.059951	0.094526	0.112980	0.127616	0.157834
min	-3.324571	-0.486894	-0.612378	-0.614466	-0.556677	-0.549263
25%	-0.312796	-0.015659	-0.030609	-0.042425	-0.054398	-0.071959
50%	0.406724	0.000792	-0.000135	0.000902	0.001554	0.009800
75%	1.161339	0.015311	0.034063	0.049184	0.056309	0.088844
max	6.277401	0.446100	0.706484	0.736165	0.688933	0.809367

Figure 9: Rudimentary quintile 2 Descriptive Stats

Quintile 2 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000
mean	0.398376	-0.000117	-0.001095	-0.003470	0.001194	-0.001346
std	1.173880	0.052858	0.077983	0.093663	0.122157	0.161286
min	-3.378045	-0.378595	-0.320004	-0.290818	-0.452630	-0.923453
25%	-0.318316	-0.017737	-0.033604	-0.048014	-0.060861	-0.081055
50%	0.353021	0.000946	-0.003947	-0.005117	-0.001219	-0.002100
75%	1.063556	0.015626	0.027824	0.034090	0.054917	0.074153
max	4.745527	0.446100	0.570480	0.454244	0.886869	0.875134

Figure 10: Less Rudimentary quintile 2 Descriptive Stats

Quintile 3 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	326.000000	326.000000	326.000000	326.000000	326.000000	326.000000
mean	0.535078	0.002502	0.004640	-0.001105	0.001164	0.001345
std	1.182658	0.042942	0.066426	0.082039	0.100950	0.161635
min	-3.706080	-0.214208	-0.224225	-0.267090	-0.357897	-0.923453
25%	-0.196259	-0.017048	-0.028362	-0.047357	-0.057809	-0.077537
50%	0.406292	0.001049	-0.001106	-0.002700	0.000439	0.002599
75%	1.145434	0.021621	0.039030	0.043862	0.061445	0.081723
max	4.809062	0.307839	0.410176	0.426202	0.433836	0.809367

Figure 11: Rudimentary quintile 3 Descriptive Stats

Quintile 3 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	482.000000	482.000000	482.000000	482.000000	482.000000	482.000000
mean	0.453687	0.000037	0.000315	0.000313	0.002470	0.003989
std	1.111879	0.054053	0.080757	0.103847	0.124095	0.147877
min	-3.216450	-0.391773	-0.415936	-0.393690	-0.475478	-0.432009
25%	-0.292602	-0.020073	-0.033047	-0.048286	-0.059946	-0.079118
50%	0.408963	0.000412	-0.002471	0.000544	0.001311	0.007613
75%	1.038950	0.017402	0.029722	0.042657	0.054521	0.074989
max	4.083637	0.363340	0.706484	0.736165	0.619469	0.735187

Figure 12: Less Rudimentary quintile 3 Descriptive Stats

The ATO mean scores increase constantly for quintiles 0-4 which is inline with expectations.

Both the rudimentary approach and the less rudimentary approach have a similar behaviour where the ATO graphs tend to be positively skewed whereas CAR0, CAR1, and CAR2 are negatively skewed and CAR3 and CAR5 make a zero centred bell curve.

In general we can say that higher and lower quintiles have stark difference of performance between the nlp methods we undertook. Our performance could have improved drastically by using FinBERT model specially trained for financial usecases.

Quintile 4 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	483.000000	483.000000	483.000000	483.000000	483.000000	483.000000
mean	0.403742	0.002209	0.004312	0.008341	0.014157	0.007329
std	1.084765	0.049182	0.078449	0.100000	0.117914	0.149512
min	-2.782408	-0.391773	-0.443378	-0.860699	-0.942553	-1.056260
25%	-0.292095	-0.017831	-0.028424	-0.037084	-0.042423	-0.069688
50%	0.331630	0.001849	0.001252	0.003716	0.005697	0.003500
75%	1.000686	0.017610	0.032726	0.049151	0.063538	0.082807
max	3.975881	0.363340	0.570480	0.450776	0.433763	0.740616

Figure 13: Rudimentary quintile 4 Descriptive Stats

Quintile 4 Descriptive Stats:

	ATO	CAR0	CAR1	CAR2	CAR3	CAR5
count	483.000000	483.000000	483.000000	483.000000	483.000000	483.000000
mean	0.294944	0.004437	0.006598	0.011710	0.014695	0.017785
std	1.166418	0.049509	0.090681	0.107077	0.120459	0.165621
min	-3.706080	-0.166610	-0.575450	-0.578227	-0.582609	-0.552741
25%	-0.500158	-0.015348	-0.030411	-0.039111	-0.045939	-0.064112
50%	0.242104	0.000378	0.003886	0.004928	0.010621	0.008607
75%	0.990477	0.017300	0.038893	0.054613	0.070936	0.092691
max	4.218503	0.524418	0.589578	0.643714	0.634667	0.992884

Figure 14: Less Rudimentary quintile 4 Descriptive Stats