BAN432 Applied Textual Data Analysis for Business and Finance

Fall 2022

Group number 8

Candidate number 25, 56 & 67

Table of contents

Task 1 – Explorative analysis

1.1 What determines newspaper coverage regarding a given firm? Please consider daily returns and stock market volume as potential explanatory variables

For this report, the word count for a specific firm determines the amount of newspaper coverage. This is because more words = more talking about the firm. Another option is using the number of articles. However, the length of articles can vary greatly and based on this we chose to use the number of words.

Based on our results in Figure 1, the larger the stock market volume is, the longer the newspaper coverage is (i.e., more words). This may be because the larger firms often have a larger following. The media pays more attention to these firms and therefore more people are invested in the firm's performance. However, the coefficient is low, which means the correlation is not strong (R^2 = 0,0006469). As for why the correlation between news coverage and return is insignificant, it is because of multicollinearity, which means stock market volume and returns are related. Another possible explanation is because stock market volume has a much larger effect on news coverage than stock returns.

```
Call:
lm(formula = log(WC) ~ abs(ret) + abs(vol), data = finaldata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7274 -0.4229  0.0511  0.4791  5.0301

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  6.338e+00  6.204e-03 1021.569  < 2e-16 ***
abs(ret)    -6.030e-02  1.595e-01   -0.378 0.705435
abs(vol)     9.293e-10  2.741e-10    3.391 0.000698 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.725 on 19738 degrees of freedom
Multiple R-squared:  0.0006469, Adjusted R-squared:  0.0005457
```

Figure 1. Regression analysis on the relationship between word count and the

1.2 When relative to the return reaction do newspapers cover the news?

We believe that the relationship between newspaper coverage and a firms performance occurs in a cyclical nature. When a change, either positive or negative, happens within the firm, this could affect the company's return. This, in turn results in the newspapers covering the news which results in stakeholders taking actions – again affecting a firms return.

From our analysis we can see the return one day before the newspaper covering is the most significant, which leads us to believe that the return reaction happens one day before the news coverage (Figure 2). In other words, this reflects the lag in news release. Articles released today are based on yesterday's information, for example, stock returns.

It is important to note that the model has a low R^2, meaning that the return may not be the most suitable explanatory variable for the news coverage.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.306411   0.007409 851.186  < 2e-16 ***
abs(ret3_before) 0.372603   0.156153   2.386   0.0170 *
abs(ret2_before) 0.331671   0.152213   2.179   0.0293 *
abs(ret1_before) 0.910370   0.127276   7.153 8.81e-13 ***
abs(ret)        -0.666585   0.170309  -3.914 9.11e-05 ***
abs(ret1_after)  0.157412   0.203979   0.772   0.4403
abs(ret2_after)  0.064201   0.214410   0.299   0.7646
abs(ret3_after)  0.054233   0.169865   0.319   0.7495
abs(ret4_after)  0.189522   0.221147   0.857   0.3915
abs(ret5_after)  0.294833   0.217193   1.357   0.1746
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7234 on 19685 degrees of freedom
Multiple R-squared:  0.005966,  Adjusted R-squared:  0.005511
F-statistic: 13.13 on 9 and 19685 DF,  p-value: < 2.2e-16
```

Figure 2. Regression on the relationship of the return reaction on the newspaper coverage.

Task 2 - Construct a sentiment dictionary

2.1 Use your insight from Task 1 to determine the right timing to measure return. Use market returns to decide what words are positive/negative.

Based on Task 1, we used returns one day before the news release. We divided the articles into two groups (positive and negative) according to whether their market returns were positive or negative.

2.2 Identify words that will capture sentiment in a newspaper context.

We selected the top 1500 documents for most negative and positive returns to create a Document Term Matrix in each group. We believe this to be enough articles to cover the majority and be specific, without removing too much vital information. We transformed all words into lower case, removed punctuations, numbers, stop words, line break marks (‘/n’), and words with less than four and more than 20 letters. We stemmed words, and selected all nouns, verbs, adjectives, and adverbs. Thus, we got two wordlists (positive and negative), which consisted of words and their frequencies. Words with low frequencies are seldomly used, and some of them are non-ASCII characters. To reduce noise, we selected words with a frequency greater than 100.

2.3 Construct a sentiment measure.

We merged negative and the positive wordlists into one, calculated the difference in the frequency of each word in the two wordlists. To assign a score to each word, we used the percentage of the frequency difference in the total word frequency. Thus, all words with a score greater than zero are positive words, and those with a score less than zero are negative words.

Our positive words were words such as ‘rose’, ‘earn’, ‘rise’, ‘gain’ and ‘profit’, which can all be associated with positive words from our judgement. Words such as ‘vaccine’ imply that stocks in the biology industry are yielding high returns due to COVID-19. The negative words such as ‘card’, ‘amex’, ‘food’, ‘service’, ‘priceline’ (an online travel agency), ‘airline’ imply that stocks in consumption, tourism and aviation industries perform poorly, yielding low returns. However, words such as ‘take’, ‘year’, and ‘said’ are not as clearly negative and may be a limitation in this report.

	word	score		word	score
1	cent	0.242125009	1	card	-0.09840445
2	stock	0.240939413	2	brand	-0.08905142
3	rose	0.196808903	3	will	-0.08588983
4	share	0.180210562	4	food	-0.08035705
5	gain	0.124092361	5	execut	-0.07877625
6	investor	0.113026800	6	advertis	-0.07021362
7	nasdaq	0.110260410	7	year	-0.05980672
8	fell	0.106308424	8	agenc	-0.05203448
9	vaccin	0.105649759	9	product	-0.05177102
10	said	0.092344740	10	offer	-0.05019022
11	earn	0.090895678	11	servic	-0.04834596
12	gamestop	0.073111741	12	program	-0.04795076
13	sale	0.072321344	13	group	-0.04505264
14	report	0.067842426	14	busi	-0.04452571
15	index	0.067578961	15	pricelin	-0.04452571
16	quarter	0.065866433	16	custom	-0.04307665
17	close	0.065734701	17	diageo	-0.04307665
18	profit	0.062704845	18	corp	-0.04268145
19	trade	0.062046180	19	amex	-0.04254972
20	expect	0.057435530	20	consum	-0.04189105
21	billion	0.054010475	21	engin	-0.04004679
22	week	0.052429681	22	presid	-0.03833426
23	maker	0.049136359	23	airlin	-0.03820253
24	advanc	0.047292099	24	unit	-0.03767560
25	ralli	0.046765168	25	offic	-0.03688520
26	thirdquart	0.045974771	26	accord	-0.03556787
27	jump	0.045447839	27	compani	-0.03490921
28	loss	0.045316106	28	make	-0.03477748
29	price	0.043998778	29	american	-0.03411881

Figure 3. List of the most positive and negative words.

Task 3 – Internal validity

3.1 Report the relationship between sentiment and return (e.g., in a regression and a plot) for firms of Group A. Do the same for the remaining firms (Group B). Are there differences in performance and why?

As can be seen in the plot in Figure 4, there is a positive relationship between sentiment and stock returns in group A. If return increases with 1 unit, news coverage will increase 3,5031%. The plot is based on the absolute values, and therefore does not go into the negatives. But the negative sentiment is affected by negative returns, and we can always get a right estimation based this model.

As for Group B, we got a higher R^2 (0,02897) and a higher coefficient (0,0487059), which means that our dictionary can explain articles in Group B even better than Group A (Figure 6). When comparing the plots in Figures 4 & 7, there is a small difference in the performance because only Group A has been used to construct the sentiment dictionary, so there may be words in there that are more frequently used in the articles in Group B. We ran this regression model several times with different random sets of Group A and Group B. The results changed slightly but are all significant, and our dictionary either fits Group A better or fits Group B better. In other words, our dictionary can explain WSJ Articles well.

Group A

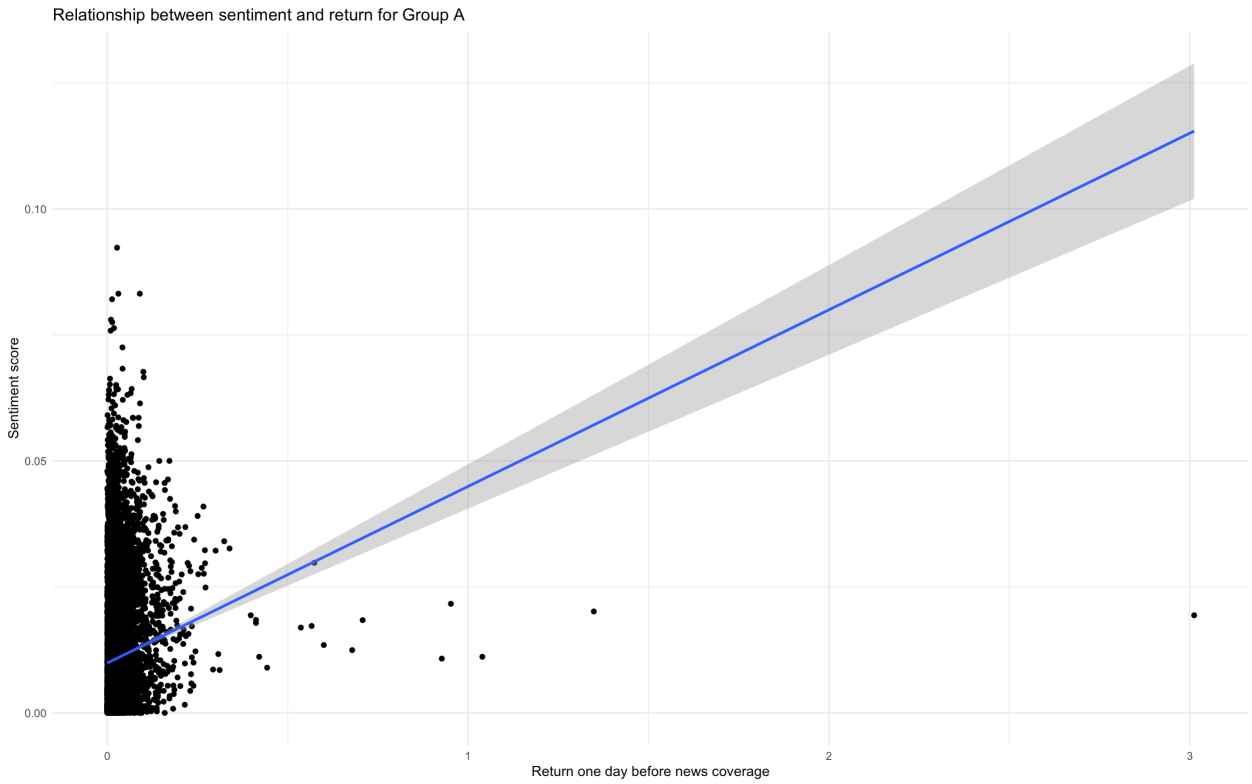

Figure 4. The relationship between sentiment and return for Group A.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0098962  0.0001307   75.73   <2e-16 ***
abs(ret1_before) 0.0350310  0.0022934   15.28   <2e-16 ***
Multiple R-squared:  0.02286,   Adjusted R-squared:  0.02277
```

Figure 5. Regression output for Group A.

Group B

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0086830  0.0001376   63.12   <2e-16 ***
abs(ret1_before) 0.0487059  0.0028558   17.05   <2e-16 ***
Multiple R-squared:  0.02897,   Adjusted R-squared:  0.02887
```

Figure 6. Regression output for Group B.

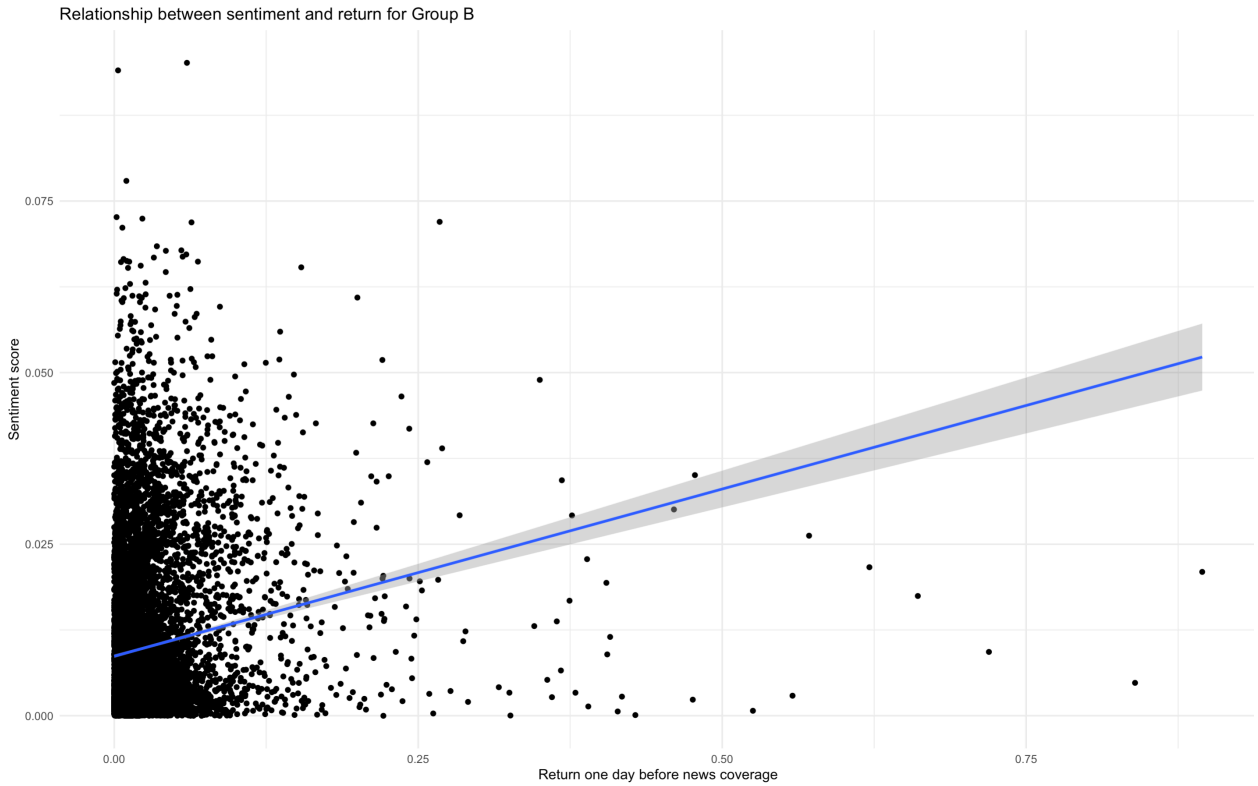

Figure 7. The relationship between sentiment and return for Group B

3.2 Split the remaining firms in Group B in firms with a lot of news articles and the ones with few articles. Make sure that the samples have approximately equal number of unique firms. Run the same regression with both sub-groups. Are there differences and why?

As can be seen in the plots in Figure 8 & 11, companies with a larger number of news articles are more closely related to our sentiment dictionary. Companies that have many articles have both higher coefficient (0,0612969) and R^2 (0,03351) than those that have few articles (coefficient = 0,0332992, R^2 = 0,0214) based on our regression model. Thus, our dictionary explains companies with many articles better than companies with few articles. This is because companies with many articles have much larger word counts, which offer a large sample for regression analysis, meaning that the smaller companies may be underrepresented.

Companies with many of articles in Group B

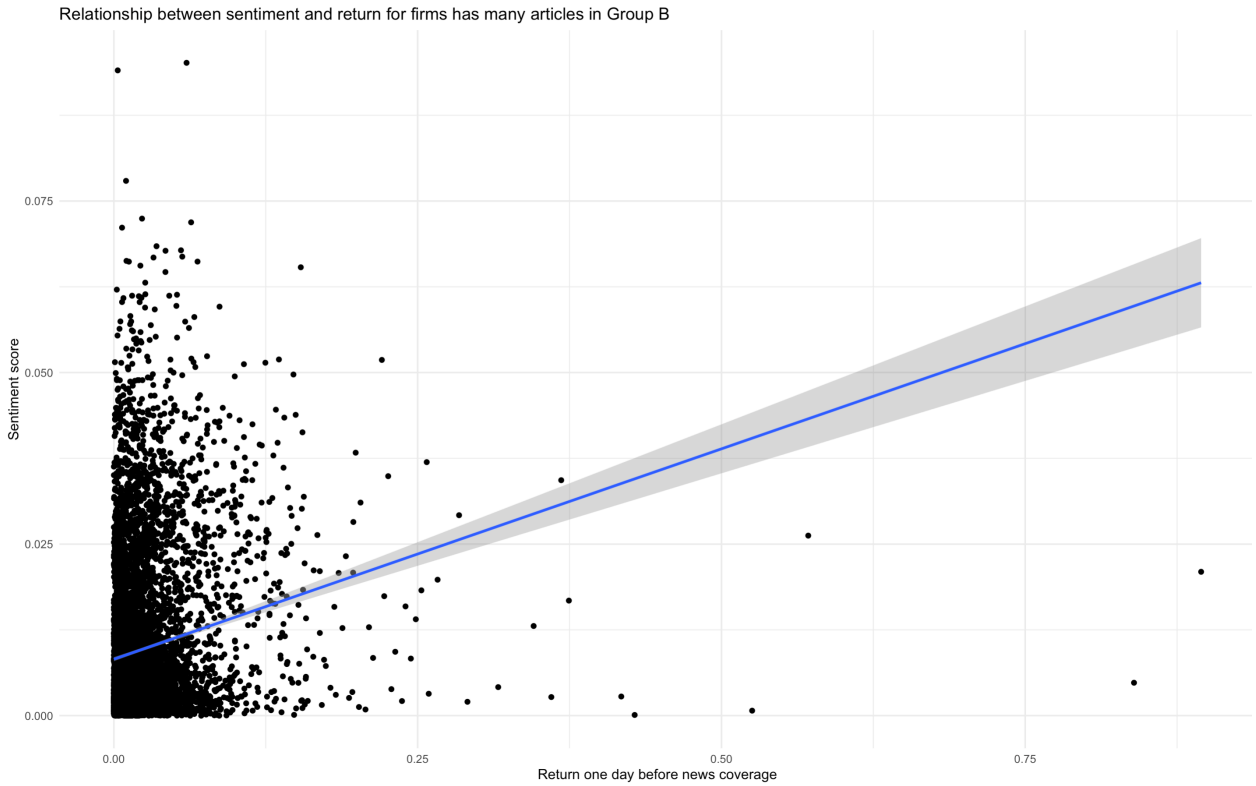

Figure 8. The relationship between sentiment and return for companies with many articles in Group B.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.0082334  0.0001565   52.62   <2e-16 ***
abs(ret1_before) 0.0612969  0.0037971   16.14   <2e-16 ***
Multiple R-squared:  0.03351,   Adjusted R-squared:  0.03338
```

Figure 9. Regression output for companies with many articles in Group B.

Companies with few articles in Group B

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.0095975  0.0002634  36.438  < 2e-16 ***
abs(ret1_before) 0.0332992  0.0042646   7.808 8.14e-15 ***
Multiple R-squared:  0.0214,    Adjusted R-squared:  0.02105
```

Figure 10. Regression output for companies with fewer articles in Group B.

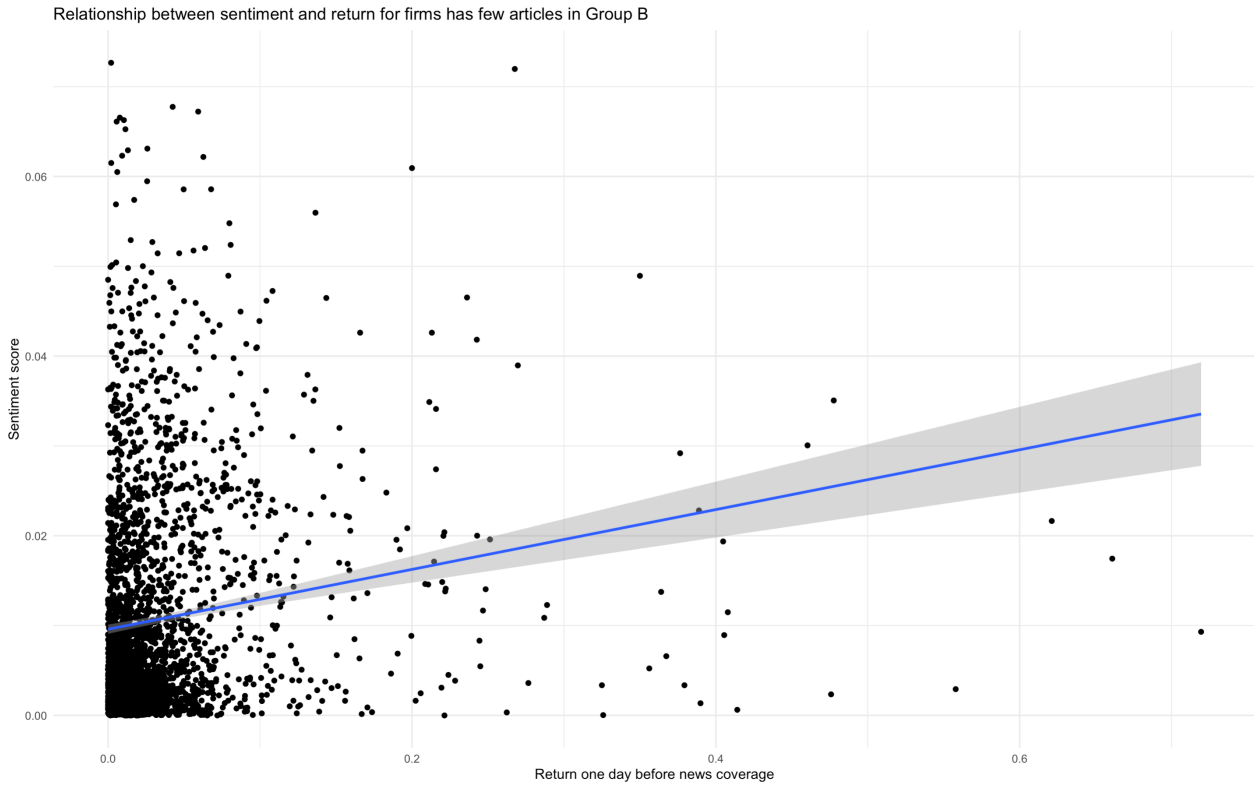

Figure 11. The relationship between sentiment and return for companies with few articles in Group B.

3.3 Change relevant parameters in the building of your sentiment dictionary and show how performance, in and out-of-sample, behaves.

We changed the number of words in the dictionary. Instead of removing the 100 least frequent words, we kept the 1000 most frequent words and ended up with a new dictionary sample. When compared to before the change (Figures 5 & 6), the explaining ability of the model on Groups A and B has decreased (Figures 13 & 15). This can be explained by the fact that less words are being taken into consideration, which in turn may exclude some articles for the lack of words.

	word	score		word	score
1	cent	0.222249506	1	card	-0.085567313
2	rose	0.172763285	2	compani	-0.063141911
3	share	0.134051280	3	product	-0.055875580
4	sale	0.110999471	4	brand	-0.046855307
5	stock	0.105737645	5	will	-0.046604744
6	profit	0.086444284	6	accord	-0.042846297
7	fell	0.085066186	7	pricelin	-0.042345171
8	quarter	0.081934147	8	peopl	-0.040591229
9	airlin	0.080054923	9	year	-0.039839539
10	earn	0.078050418	10	motorola	-0.037584471
11	gain	0.077799855	11	offer	-0.036331655
12	nasdaq	0.075670068	12	american	-0.035579966
13	revenu	0.072287466	13	program	-0.035204121
14	vaccin	0.070784087	14	drug	-0.035078840
15	airway	0.068278456	15	amex	-0.034702995
16	report	0.062515504	16	make	-0.032573208
17	xerox	0.059508746	17	fund	-0.031570956
18	billion	0.057378959	18	time	-0.031320393
19	expect	0.057253678	19	take	-0.030318140
20	said	0.056251425	20	visa	-0.029065324
21	loss	0.050488473	21	like	-0.028940043
22	sprint	0.048233405	22	presid	-0.028564198
23	post	0.045226647	23	onlin	-0.028564198
24	thirdquart	0.044600239	24	manag	-0.028438916
25	million	0.040591229	25	advertis	-0.028313635
26	result	0.040215384	26	custom	-0.028188353
27	carrier	0.039087850	27	unilev	-0.026810256
28	moderna	0.038962568	28	part	-0.026559693
29	maker	0.038712005	29	food	-0.026309130

Figure 12. List of the most positive and negative words.

Group A

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.0100923  0.0001332   75.74   <2e-16 ***
abs(ret1_before) 0.0363369  0.0023384   15.54   <2e-16 ***
Multiple R-squared:  0.02364,   Adjusted R-squared:  0.02355
```

Figure 13. Regression output for Group A when a parameter is changed.

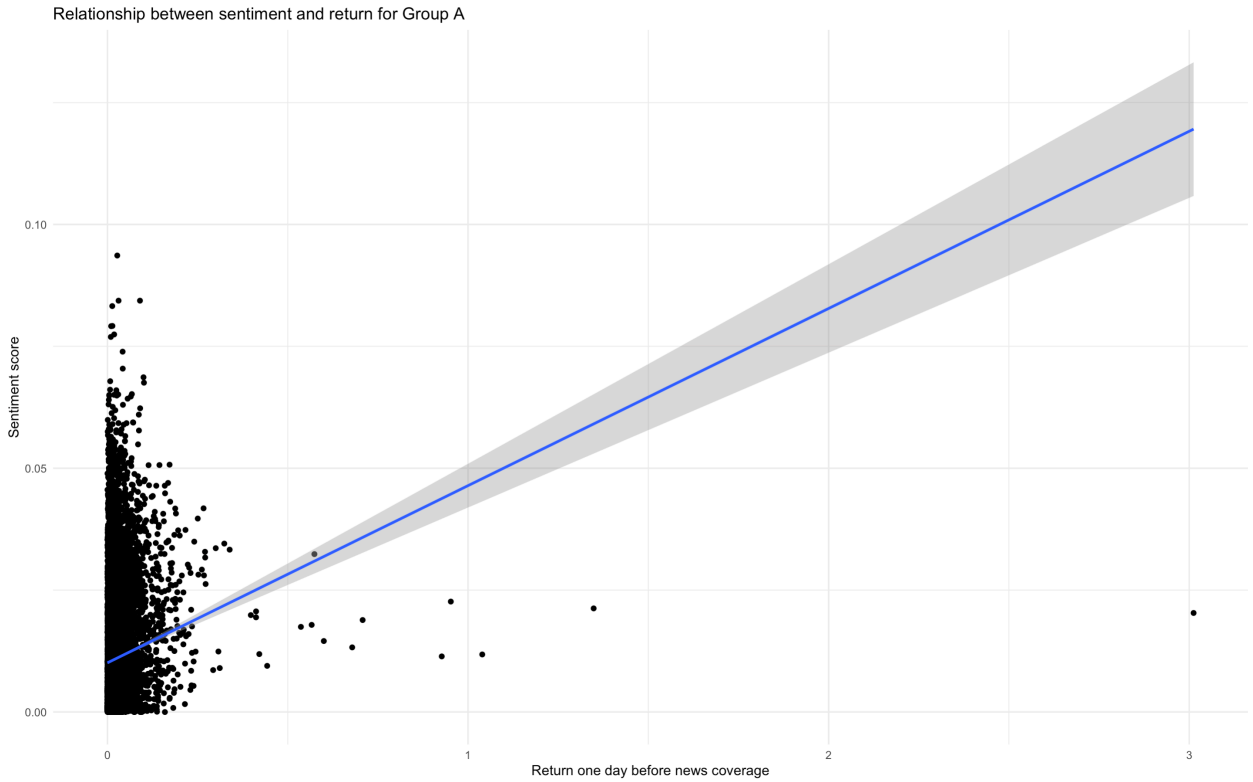

Figure 14. The relationship between sentiment and return for Group A when a parameter is changed.

Group B

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.008852   0.000140   63.22   <2e-16 ***
abs(ret1_before) 0.049822   0.002907   17.14   <2e-16 ***
Multiple R-squared:  0.02924,   Adjusted R-squared:  0.02914
```

Figure 15. Regression output for Group B when a parameter is changed.

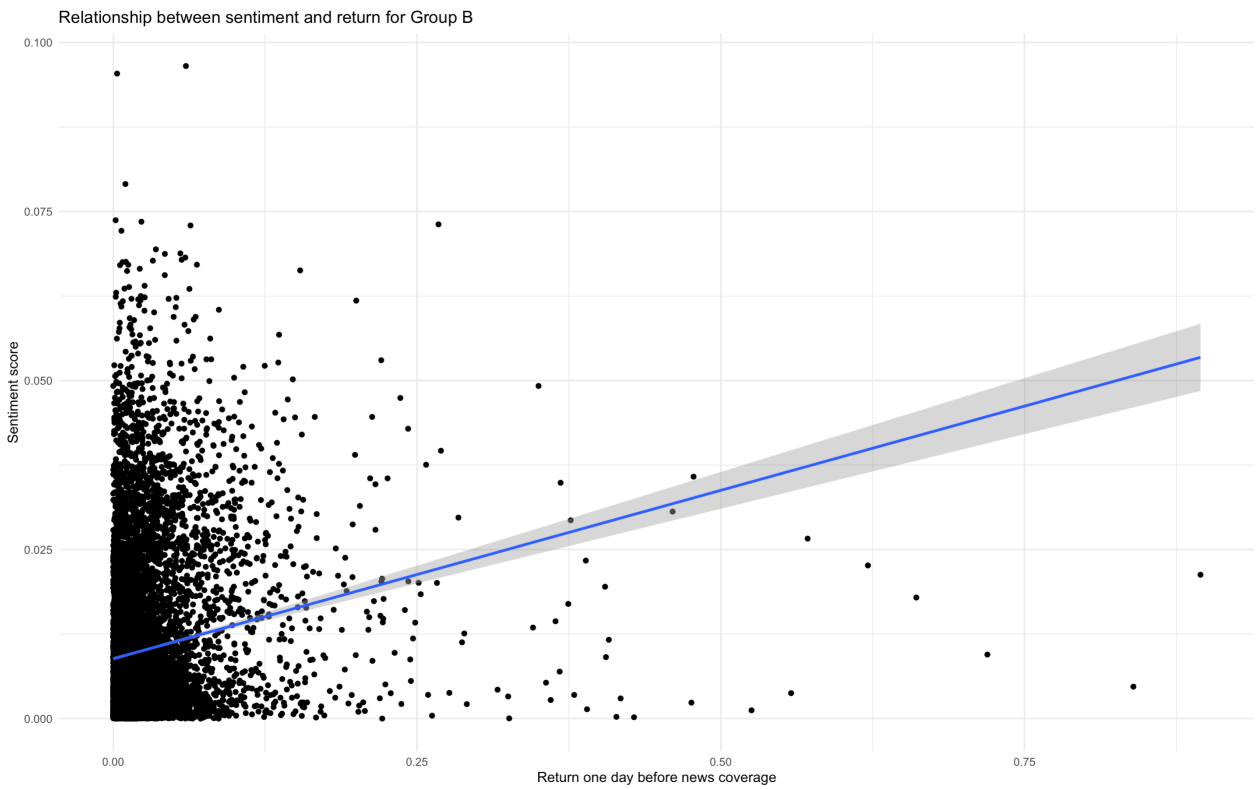

Figure 16. The relationship between sentiment and return for Group B when a parameter is changed.

Furthermore, when the parameter for word frequency changed from higher than 100 to 1000 words with highest frequency, we would tell that the coefficient for Group B increased from 0,0487059 to 0,049822, and the R-square changed from 0,0288 to 0,0291. Although, in this instance, there was not a very large change, this still demonstrates that 1000 words with highest frequency have better efficiency than words with frequency higher than 100. We believe that having too many words might cause more noise and bring less significant results.

Besides, we also conducted the analysis with different numbers of documents (500, 1000, 2000) for most negative and positive returns to build the Document Term Matrix. However, word lists we got were not ideal enough to construct a sentiment dictionary and to further our analysis.

Based on these analyses we cannot conclude that our internal validity is good enough as R squared is very low. It may be a case of lacking construct validity, as we cannot be sure that our dictionary accurately measures the right negative and positive words.

Task 4 – External validity

4.1 Use a corpus of earnings calls and apply your sentiment dictionary. Investigate in a regression if return and sentiment correlate. In addition to considering the full sample, apply the same split in Group A and B. Does it do equally well as in Task 3?

As can be seen from these regression analyses, the external validity is not very high, and the results are not significant. We are not able to generalize this dictionary as it produces inefficient results when used on other, external corpora, such as the corpus of earning calls. One possible explanation is that news articles have a different structure from earning calls. In our analysis, we used Q&A sessions to build the corpus of earning calls. The wording of news articles is much more cautious, professional, and formal, and the wording of earning calls, especially Q&A sessions, is more casual and informal. Such differences may lead to different semantics and emotions, which have a great effect on sentiment measurement of these texts.

Earning calls - The full sample

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.0022529  0.0001468  15.345   <2e-16 ***
abs(ret1_before) -0.0019089  0.0037209  -0.513    0.608
Multiple R-squared:  0.0008296, Adjusted R-squared:  -0.002322
```

Figure 17. Regression output for the full sample.

Earning calls - Group A

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.002825   0.000280  10.091   <2e-16 ***
abs(ret1_before) -0.009671   0.006350  -1.523     0.13
Multiple R-squared:  0.01714,   Adjusted R-squared:  0.009752
```

Figure 18. Regression output for Group A in Earning calls.

Earning calls - Group B

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.0018074  0.0001425   12.68   <2e-16 ***
abs(ret1_before) 0.0058717  0.0039947    1.47    0.143
Multiple R-squared:  0.01173,   Adjusted R-squared:  0.006302
```

Figure 19. Regression output for Group B in Earning calls.

Earning Calls - Group B with many articles

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.0017732  0.0001642  10.799   <2e-16 ***
abs(ret1_before) 0.0069044  0.0049543   1.394    0.166
Multiple R-squared:  0.0158,    Adjusted R-squared:  0.007663
```

Figure 20. Regression output for Group B with many articles in Earning calls.

Earning Calls - Group B with few articles

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.0023227  0.0003205   7.248 2.74e-09 ***
abs(ret1_before) -0.0021929  0.0075609  -0.290    0.773
Multiple R-squared:  0.001714,  Adjusted R-squared:  -0.01866
```

Figure 21. Regression output for Group B with few articles in Earning calls.