BAN432
Applied Textual Data Analysis for Business and Finance

Exam fall 2022

Abstract

We construct a sentiment dictionary using newspaper articles from the Wall Street Journal (WSJ) and contemporaneous stock returns to determine which words are positive or negative in a financial context. Applying a Multinomial Inverse Regression model (MNIR) by Taddy (2013), we find that our constructed sentiment dictionary for bigrams relate to returns. We also create an *alternative* dictionary to compare with the MNIR – which is based on simpler tools than machine learning. We find, using MNIR, that bigrams have internal validity and weak external validity.

Data Wrangling

The data provided is unstructured. In order to analyze the articles, we need to convert the files into a structured format. Since we need stock market data for all tasks, we only load articles for companies that are present in the stock returns data. By inspecting the html-files, we find that each article is stored in a html table-format. This enables us to load all the articles in one list, where each element is a table with rows that helps us identify the different parts of each article (publishing date, word count, etc.).

1.0 Explorative Analysis

1.1 What determines newspaper coverage regarding a given firm?

We define newspaper coverage as the number of articles published about a given firm. We will therefore ignore the content and the sentiment of the articles in this part of the paper. Based on the given data, we hypothesize that company size, trading volume and stock price movements can be explanatory variables for newspaper coverage. For the explorative analysis, we divide the total number of articles related to each company published in the timeframe by the number of years of data we have related to that company. This enables us to compare the number of articles published for companies with different years of data.

To determine company size, we use the median market cap for each company in the period. The market cap is computed by multiplying the price[1] by the shares outstanding. Figure 3 shows that there is a positive relationship between company size and how often a company is covered in the WSJ. There are however very few observations with a market cap above $20 billion, which means that more data is needed to accurately assess this relationship.

Due to the large number of observations, daily returns are approximately normally distributed for most firms. Thus, the median and mean will be close to zero for most of the companies, making it impractical for interpretation. To determine whether WSJ writes more about companies with large price fluctuations, we use average absolute returns for each firm and plot it in the same manner as for market cap. Figure 4 shows that there is a slightly negative correlation between these variables for the companies in the dataset. Similarly, to company size, there are a few observations with very large average absolute returns. Even when ignoring these, there does not seem to be a strong relationship between these variables.

Unusually high or low levels of negative sentiment about a company are normally related with increases in trading volume for its stock (Tetlock, 2007, p. 1143). Because the daily trading volume, measured in number of stocks traded, is highly dependent on its price and shares outstanding we decide to use daily volume in USD for this analysis. This is computed by

[1] Some of the observations have negative prices, indicating that the price is given by the bid-ask average, which occurs when there is no closing price for the day (CRSP, n.d). We therefore use the absolute value of the price variable.

multiplying the price by the daily volume. Figure 5 shows that there is a weak positive correlation between these variables, implying that the WSJ writes more articles about companies with larger trading volumes.

WSJ appears to write more about companies with a high market cap and high trading volume. This is reasonable, as there usually is more to write about larger companies compared to smaller ones. Somewhat surprising, stock price movements do not have a positive correlation with the number of articles WSJ writes about the company in a year. It is important to emphasize that the purpose of these charts is to highlight broad trends in the dataset rather than to assert statistically significant relationships.

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

Even though stock price movements do not seem to influence how often WSJ writes about a company, it is interesting to study the timing of articles relative to stock price movements. We use average absolute returns to compare the price movement of a stock in the days before, on and after the day an article is published. Figure 6 shows that the price movement is larger the day before and on the day an article is published, indicating that the WSJ writes about a company after its stock has had a large movement. This likely occurs because stock related news articles lag earnings calls/reports and other news driving the stock price. The results are somewhat different from the findings by Garcia, Hu, and Rohrer (2022). This is likely due to our sample being different and containing stocks that have had abnormal stock price movements. For instance, the dataset contains multiple “meme-stocks”, such as GME, AMC and BBBY (Hayes, 2022).

2.0 Constructing a Sentiment Dictionary

2.1 Data Cleaning

Before constructing the sentiment dictionary, we need to clean the data. The cleaning steps are:

1. Removing HTML-tags, URL’s, digits, and punctuations.
2. Removing all stop words, except those containing important context for the sentiment analysis[2].
3. Removing ending sentence of articles starting with “License this article…” and “Subscribe to WSJ…”, where applicable.
4. Removing single character words and words with more than 20 letters.
5. Converting letters to lower case.
6. Removing excess whitespace.

[2] Stop words that we keep are *cannot, can't, above, under, over, hasn't, wouldn't, below, up, down, further, don't,* and *isn't.*

2.2 Multinomial Inverse Regression Model

To train the MNIR model we create a random training sample which we call Group A and a random hold out sample which we call Group B. This will allow us to test for internal validity. We only consider the body of the articles, classified as TD in the html-files, and ignore the preamble as it summarizes the body.

We use unigrams and bigrams in our model. Using a corpus of *n* documents based on tokens, we construct a document term matrix (dtm) of all documents (both A and B sample) where we keep count of the frequency of each term in each of the documents in our corpus. We represent the dtm as a sparse matrix, as this is more computationally efficient, reduces memory usage and lets us construct the multinomial inverse regression.

From the above dtm we keep all bigrams that occur in our training sample and remove those in our hold out sample. Further, we apply multiple parameters to ensure consistency across the dtm. First, we construct an Inverse Document Frequency (IDF) score for each term – this is a measure of whether a term is common or rare in each document (Nettleton, 2014, p. 175). Using the IDF score we penalize terms that are rare across all documents. Further, we consider only the top 20 terms in each document and then remove firm-specific terms to reduce overfitting. I.e., some firms have a high number of articles relative to other firms, as shown in Figure 2. This means that firm-specific terms can cause noise in our dictionary, reducing the validity[3].

The predicted attribute – contemporaneous returns – is computed as the cumulative return over a three-day event-window based on the results we show in Figure 6. This is a similar approach to Garcia, Hu, and Rohrer (2022), and ensures that we capture the stock price movements related to the article, as this usually occurs the day before or on the same day as the article is published.

The top constructed *n*-grams are shown in Table 4 (bigrams) and Table 5 (unigrams). Amongst the top positive terms for bigrams are *shares rose*, *price target*, *Nasdaq rose*, *fourth-quarter profit*, and *income rose*. And amongst the top negative terms are *fell points*, *shares fell*, *52-week low*, *closed down*, and *stock fell*. Notice the conjugation of the verbs in the bigrams. This underpins our point in task 1 and Figure 6. I.e., news article coverage is determined by past moves in the stock price.

If we examine the top unigrams we find similar results, with top positive terms such as *rose*, *accelerated*, and *climbed*. And top negative terms are *fell*, *52-week*, and *below*. Here we see an important flaw of unigrams, as they lack the context that bigrams have. For example, *52-week* can be both *52-week low* and *52-week high*. And as we see in Table 4 the correct context is *52-week low*.

[3] For example, bigrams such as *cable operator* and *air group* appear frequently in our sample. However, these bigrams are only mentioned in articles related to Time Warner Cable and American Airlines.

2.3 Intuitive approach

To compare the dictionary output of the MNIR model, we decided to create an alternative model using a more intuitive approach. We used a TF–IDF (Term Frequency – Inverse Document Frequency) score to filter out the most relevant terms just like we did with the MNIR input. Then we computed a sentiment measure by calculating the average return (contemporaneous return) for each term across all the documents.

The top 20 positive and negative terms are presented in Table 4 (bigrams) and 5 (unigrams). When comparing the top terms of the dictionaries of bigrams, we find some interesting similarities. In both dictionaries of bigrams, we find the negative terms *shares fell* and *shares down*, and the positive terms *shares rose* and *shares up*. The top terms of the dictionaries of unigrams does not share the same similarities, but both negative dictionaries contain the term *fell*.

Note that we do *not* proceed with this approach in our evaluation of internal and external validity and use the MNIR as our main model.

3.0 Internal validity

In this section we employ the uni- and bigrams constructed in section 2.2 to test for internal validity. We report the results in Table 9 and find that sentiment for both bigrams and unigrams have a significant relationship with stock market returns. However, bigrams are a better predictor of returns, with an R^2 of 3.4 percent for Group A (training sample), and an R^2 of 0.3 percent for Group B. Not surprisingly, unigrams have less predictive power for the reasons mentioned in section 2.2. Further, we find in Table 8 higher t-values for Group A than Group B which implies that the model overfits to group A.

Based on the skewed distribution of articles across firms (Figure 2) we run additional tests on Group B. We split the firms in the hold out sample into two categories – one with firms that have a high number of articles and one with firms that have a low number of articles. As shown in Table 10, we find that the firm in the high frequency bucket has higher explorative power. This is likely because our training sample (which consists of random firms) has more firms with a high number of articles (Group A consists of more than 50 per cent of the articles we use).

Lastly, we look at how changes in relevant parameters affect performance (sensitivity) in Group A and B for bigrams. As we have more than ten unique parameters on our model, we only consider the parameters shown in Figure 9, 10, 11, and 12. The mentioned figures illustrates how (and discusses) how our model behaves in relation to these parameters.

4.0 External validity

To test for external validity, we apply our dictionary on a corpus of earnings calls transcripts. We run a regression to evaluate whether there is a correlation between the contemporaneous returns and sentiment in the earnings calls. The results presented in Table 11 and 12 show that there is no significant correlation between returns and sentiment scores for unigrams because the coefficients are not significantly different from zero. We also find that firms with a high frequency of transcripts have a higher correlation between returns and sentiment scores than firms with a low frequency of transcripts because of a higher number of observations. To conclude, we find no external validity for unigrams. However, we find some (weak) external validity for bigrams.

In contrast to standardized forms of texts, news articles are highly dependent on the time they are in. For example, an earnings report or call talk about a limited subject – i.e., financial statements, which directly impact the share price. However, news articles often consist of time sensitive and systematic subjects, such as Covid-19 ('20 - '21), the GFC ('08 - 09') or retail frenzies ('21). This implies that a sentiment dictionary for newspaper articles – in contrast to earnings reports/calls – might include time-relevant words such as *vaccine* or *Lehman Brothers.* Based on the above discussions, we agree with the claims of Garcia, Hu, and Rohrer (2022), that earnings calls are a better event for performing this type of supervised learning algorithm.

5.0 Bibliography

CRSP. (n.d). *Data Definitions - B*. Retrieved November 17, 2022, from Center for Research in Security Prices (CRSP): https://www.crsp.org/products/documentation/data-definitions-b

Garcia, D., Hu, X., & Rohrer, M. (2022, August). The Colour of Finance Words. *SSRN*.

Hayes, A. (2022, September 12). *What Are Meme Stocks, and Are They Real Investments?* Retrieved November 18, 2022, from Investopedia: https://www.investopedia.com/meme-stock-5206762

Nettleton, D. (2014). Chapter 11 - Text Analysis. In *Commercial Data Mining* (pp. 171-179). Morgan Kaufmann.

Taddy, M., 2013, "Multinomial inverse regression for text analysis," Journal of the American

Statistical Association, 108(503), 755–770.

Tetlock, P. (2007, June). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *THE JOURNAL OF FINANCE*, 1139-1168.

6.0 Tables

Table 1: Summary Statistics for WSJ after Cleaning

Wall Street Journal (WSJ)	
Start	03.01.2000
End	31.12.2021
Observations	30,352
Unique firms	91
Average words per article	332
Average articles per firm	334

Table 2: Summary Statistics for Earnings Calls Transcripts after Cleaning

Earnings Call Transcripts	
Start	29.11.2005
End	12.10.2020
Observations	333
Unique firms	31
Average words per article	4,674
Average articles per firm	11

Table 3: MNIR – filters used in preprocessing

Filters	Base value
Absolute contemporaneous returns	30%
Max firms that one news article concerns	25
Stop words kept	"above", "under", "over", etc.[4]
Max IDF	5
Top # of terms in each document	20
Minimum number of firms that regard one term	30 for bigrams and 60 for unigrams
Terms in bigram dictionary	150
Terms in unigram dictionary	75

[4] *cannot, can't, above, under, over, hasn't, wouldn't, below, up, down, further, don't,* and *isn't.*

Table 4: Top bigrams using MNIR

The Table shows top 20 positive and negative bigrams sorted by loading factor related to returns. Higher loadings indicate that articles using these bigrams tend to have high returns – measured as buy-and-hold returns over a three-day event window. Lower articles that contain low loadings bigrams tend to have negative returns. The loadings are calculated on initial 30,352 articles[5] from 91 unique firms in the period 2000-2022 that are tokenized to bigrams and filtered.

Negative		Positive	
Bigram	**Loadings**	**Bigram**	**Loadings**
fell points	-8.00	price target	6.51
stock fell	-6.58	shares rose	6.49
52-week low	-6.09	nasdaq rose	6.12
closed down	-5.92	fiscal fourth-quarter	5.80
shares fell	-5.70	fourth-quarter profit	5.53
shares down	-5.64	income rose	5.08
declined cents	-4.95	closed up	4.75
third-quarter results	-4.54	coming months	4.72
said also	-4.43	shares up	4.69
financial markets	-4.35	down million	4.64
plans sell	-4.25	million people	4.53
earlier week	-4.24	quarter million	4.14
under terms	-3.94	private-equity firms	4.00
charlotte n.c	-3.91	stock up	3.98
one year	-3.67	stock rose	3.98
lost cents	-3.63	u.s economy	3.91
down cents	-3.60	advanced cents	3.91
one big	-3.57	global economic	3.88
person close	-3.51	went public	3.83
year also	-3.49	job cuts	3.67

[5] We apply various filters on these articles to reduce overfitting and ensure consistency.

Table 5: Top 20 unigrams using MNIR.

We consider the top 20 positive and negative unigrams sorted by loading factor related to returns. Note that we have a lot of the same tokens as in Table 4. However, the terms give less context.

Negative		Positive	
Unigram	**Loadings**	**Unigram**	**Loadings**
fell	-3.64	rose	3.06
52-week	-3.06	accelerated	1.87
30-year	-2.81	added	1.77
accepted	-1.80	a.m	1.55
absolutely	-1.73	climbed	1.41
1970s	-1.62	character	1.36
although	-1.62	agreement	1.35
below	-1.61	income	1.31
information	-1.48	stake	1.29
air	-1.34	euros	1.15
concerns	-1.29	thursday	1.09
down	-1.18	energy	1.07
dropped	-1.10	access	1.04
close	-1.07	calls	1.00
agency	-1.04	began	0.98
ability	-1.04	apple	0.97
capital	-0.94	amid	0.95
stocks	-0.88	fourth-quarter	0.93
almost	-0.87	offer	0.92
loss	-0.85	ahead	0.91

Table 6: Top 20 bigrams using alternative model

Negative		Positive	
Bigram	**Loading**	**Bigram**	**Loading**
shares fell	-5.10	shares rose	4.59
shares down	-5.10	shares up	3.99
down cents	-3.18	said monday	3.59
fell cents	-2.44	market value	3.55
capital one	-2.40	stock price	3.17
earlier period	-2.30	companies like	3.14
sales fell	-2.17	market capitalization	2.39
said no	-2.09	fiscal fourth	2.23
fell million	-1.75	said wednesday	2.19
quarter year	-1.71	asset management	1.94
no longer	-1.52	warner cable	1.84
analysts said	-1.46	financial crisis	1.75
credit card	-1.42	las vegas	1.73
higher prices	-1.40	million shares	1.72
reached comment	-1.37	recent months	1.70
quarter sales	-1.36	matter said	1.68
also said	-1.33	no one	1.63
morgan stanley	-1.30	rose cents	1.62
cents cents	-1.27	high profile	1.60
company reported	-1.26	hedge fund	1.56

Table 7: Top 20 unigrams using alternative model

Negative		Positive	
Unigram	**Loading**	**Unigram**	**Loading**
trades	-4.84	soared	7.32
slid	-4.26	traders	6.15
fell	-3.36	pandemic	5.85
restrictions	-3.17	coffee	5.17
amd	-3.12	usa	4.76
tumbled	-3.11	user	4.46
check	-2.98	individual	4.34
sec	-2.85	eye	4.14
warned	-2.82	jumped	4.03
wpp	-2.78	surged	3.96
requirements	-2.64	posting	3.94
lowest	-2.63	community	3.87
suggests	-2.62	park	3.79
edge	-2.58	google	3.68
read	-2.52	driven	3.65
acknowledged	-2.50	optimistic	3.57
declines	-2.49	pushing	3.56
commodities	-2.44	gained	3.50
commodity	-2.32	options	3.49
giants	-2.30	users	3.46

Table 8: Overfitting

To test the relationship between returns and bigrams we create the following regression, where we test whether the sentiment scores are related to the firm's return at time t:

$$R_{it} = \beta_0 + \beta z_{it} + \beta m_{it} + \varepsilon_{it}$$

Where i is the articles; t is the publishing date; R_{it} is the firms buy-and-hold returns over a three-day window around the event; z_{it} is the SR-score for each article; and m_{it} is a control variable for the length of the document. *Group B* is our hold out sample and *Group A* is our training sample. As the Table shows the training sample has a higher t-value than the hold out sample meaning the model overfits.

	Dependent variable:			
	Returns			
	Group B (Bi)	Group A (Bi)	Group B (Uni)	Group A (Uni)
	(1)	(2)	(3)	(4)
z	0.005^{***}	0.013^{***}	0.002^{***}	0.004^{***}
	t = 7.471	t = 25.003	t = 3.221	t = 6.357
m	-0.00003	-0.0001^{***}	0.00001	0.00000
	t = -1.067	t = -2.649	t = 1.203	t = 0.192
Constant	0.001	0.002^{***}	-0.001	0.0002
	t = 0.541	t = 2.651	t = -1.147	t = 0.174
Observations	10,550	12,546	10,550	12,546
Adjusted R^2	0.005	0.047	0.001	0.003
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 9: Internal validity

In this Table we test the model for internal validity using the training sample (Group A) and hold out sample (Group B) with the following regression:

$$R_{it} = \beta_0 + \beta_1 Positive_{it} + \beta_2 Negative_{it} + Log(WC) + \varepsilon_{it}$$

Where i is the articles; t is the publishing date; R_{it} is the firms buy-and-hold returns over a three-day window around the event; $Positive_{it}$ is the positive sentiment score; $Negative_{it}$ is the negative sentiment score; and $Log(WC)$ is a control variable for the word count of the document. We have 150 positive and negative terms in our dictionary. All sentiment measures are standardized by its mean and standard deviation. We find that bigrams perform best out of the two groups both on training data and hold out data. And that unigrams have very weak predictive power. Note that we do not use robust standard errors in this regression.

	Dependent variable:			
	Returns			
	Group B (Bi)	Group A (Bi)	Group B (Uni)	Group A (Uni)
	(1)	(2)	(3)	(4)
Positive	0.002^{***}	0.008^{***}	0.004	-0.002
	t = 3.315	t = 14.099	t = 1.639	t = -0.812
Negative	-0.002^{***}	-0.007^{***}	0.002	-0.006^{***}
	t = -3.891	t = -12.461	t = 0.737	t = -2.754
Log(Word Count)	0.001	-0.001	0.002^{**}	0.0004
	t = 0.940	t = -0.715	t = 1.985	t = 0.472
Observations	9,700	11,450	10,543	12,534
Adjusted R^2	0.003	0.034	0.001	0.004
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 10: Internal validity – Group B split by news frequency

In this Table we split the firms in Group B (hold out sample) by news article frequency. We split the firms into two groups – with low frequency and high frequency of news articles. We define the low frequency firms as firms that have less or equal to 275 news articles and high frequency firms as firms that have more than 275 articles (see Figure 2). This results in 46 firms in the high frequency group and 45 firms in the low frequency group. As the Table displays, firms with a high frequency of news articles have a higher predictive power than low frequency firms. This occurs, even though we punish high frequency terms that appear for few firms.

	Dependent variable:			
	Returns			
	High Freq. (Bi)	Low Freq. (Bi)	High Freq. (Uni)	Low Freq. (Uni)
	(1)	(2)	(3)	(4)
Positive	0.002^{***}	0.002	0.001	0.0002
	t = 3.254	t = 1.104	t = 1.340	t = 0.098
Negative	-0.002^{***}	-0.002	-0.001^{*}	-0.003
	t = -3.600	t = -1.478	t = -1.721	t = -1.619
Log(Word Count)	0.002	-0.001	0.002^{**}	0.0002
	t = 1.412	t = -0.407	t = 2.191	t = 0.087
Observations	7,513	2,187	8,176	2,367
Adjusted R^2	0.003	0.0004	0.001	0.0004
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 11: External Validity

The sentiment dictionary created based on WSJ articles are applied to earnings calls transcripts to test external validity. The regression model follows that of Table 9. This Table proves the initial point that we made in chapter 2.2. I.e., unigrams lack context which means that unigrams with negative sentiment scores are fitted to words that are used in a positive context. This means (for unigrams) that the positive and negative sentiment scores are uncorrelated with the dependent variable because the coefficients are not significantly different from zero.

	Dependent variable:	
	Returns	
	Bigrams	Unigrams
	(1)	(2)
Positive	0.004	-0.004
	t = 0.751	t = -0.688
Negative	-0.012^{**}	0.007
	t = -2.272	t = 1.339
Log(Word Count)	-0.038^{*}	-0.035^{*}
	t = -1.945	t = -1.689
Observations	333	333
Adjusted R^2	0.013	0.005
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 12: External validity – earning calls split by news frequency

We split the firms into two groups – one with high frequency of earning calls and one with low frequency of earning calls. We have 31 firms in total and 15 firms in our high frequency group and 16 firms in our low frequency group. Not surprisingly we see the same results that we discuss in Table 10 for unigrams – there is no significant correlation between the dependent and independent variables. The main focus of this table is the bigrams, and the main point is the same that we show in Table 9 for internal validity; that firms with a high frequency of transcripts have a higher predictive power and correlation than low frequency firms. This is because high frequency firms have more observations.

	Dependent variable:			
	Returns			
	High Freq. (Bi)	Low Freq. (Bi)	High Freq. (Uni)	Low Freq. (Uni)
	(1)	(2)	(3)	(4)
Positive	0.016^{**}	-0.007	0.00000	-0.008
	t = 2.568	t = -0.770	t = 0.0001	t = -0.625
Negative	-0.018^{***}	-0.008	-0.002	0.017^{*}
	t = -2.900	t = -0.527	t = -0.308	t = 1.684
Log(Word Count)	-0.042^{*}	-0.062	-0.018	-0.091^{*}
	t = -1.900	t = -1.169	t = -0.821	t = -1.913
Observations	234	99	234	99
Adjusted R^2	0.040	0.003	-0.009	0.035
Note:				*p<0.1; **p<0.05; ***p<0.01

7.0 Figures

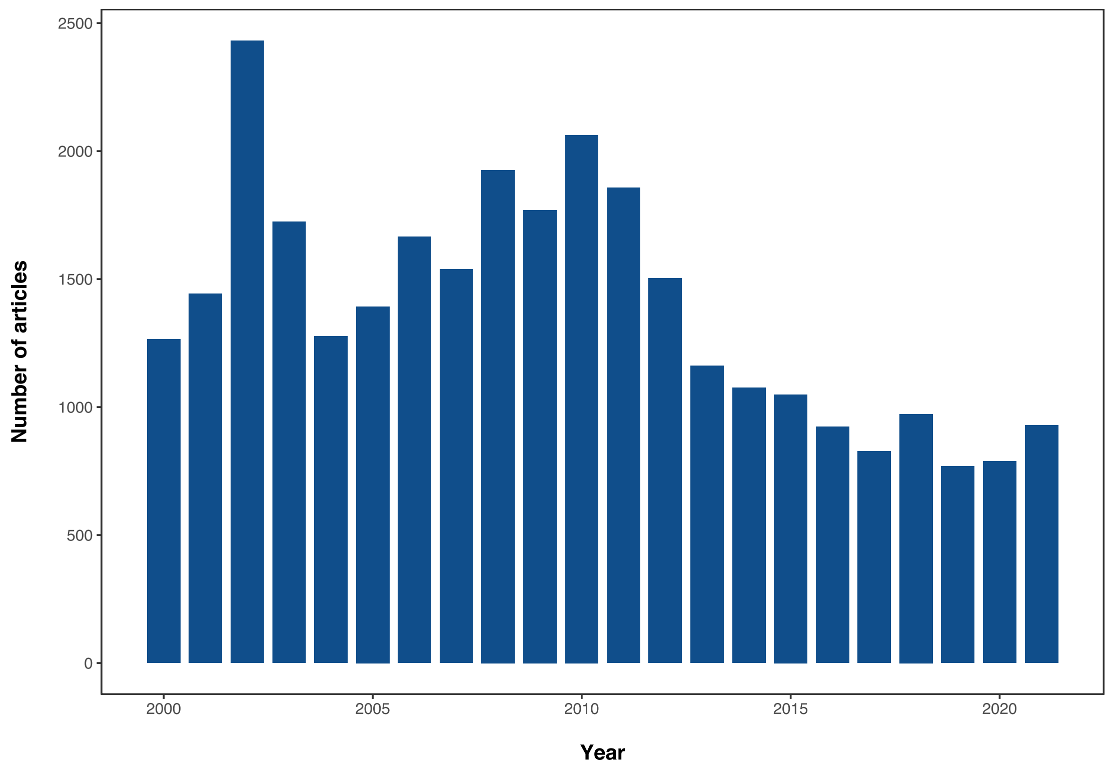

Figure 1: Articles by year

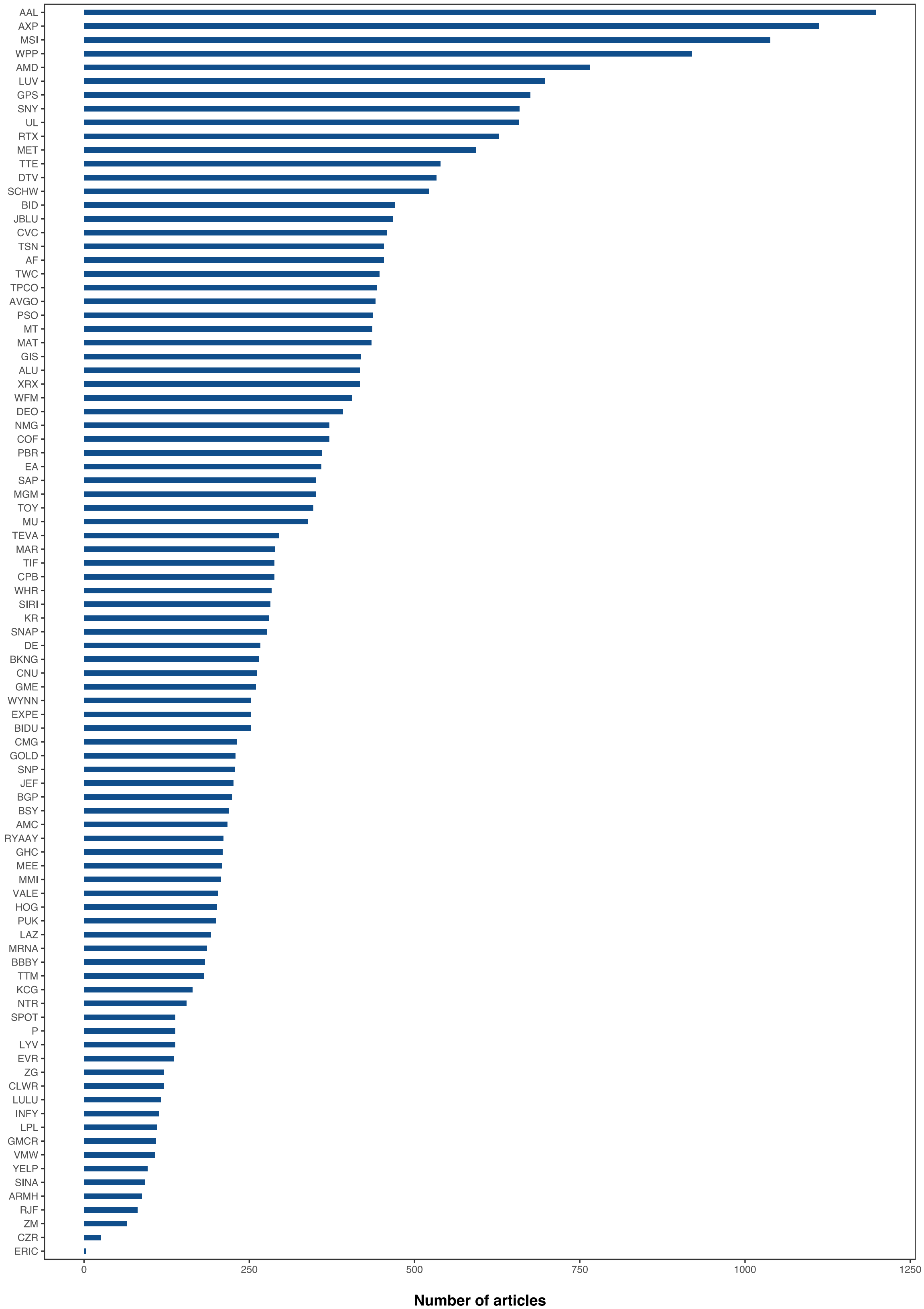

Figure 2: News article frequency by firm

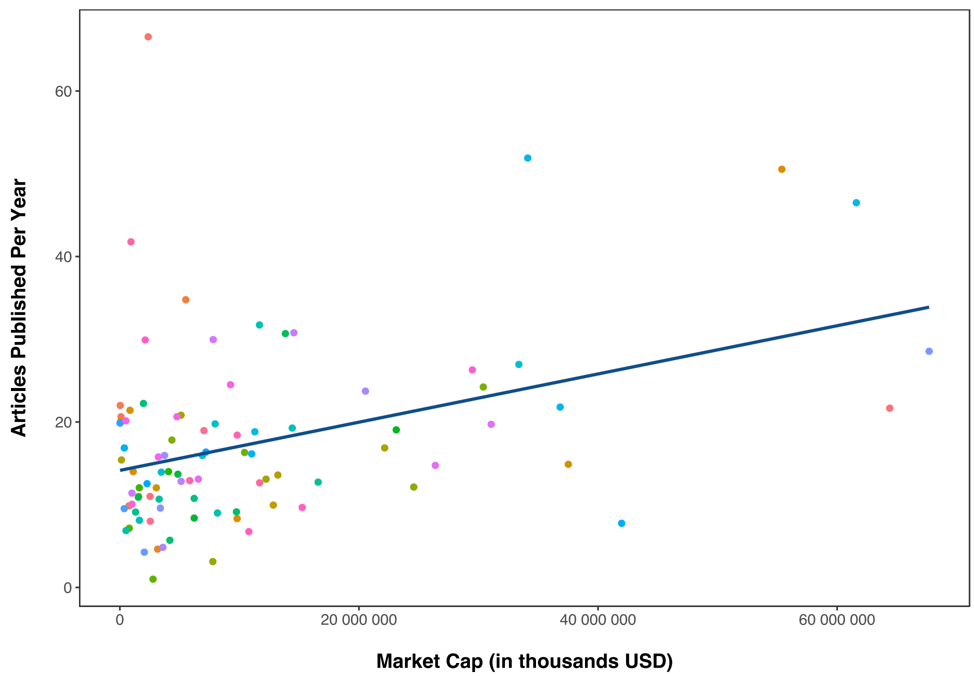

Figure 3: Relationship between Median Market Cap and Articles Published per Year

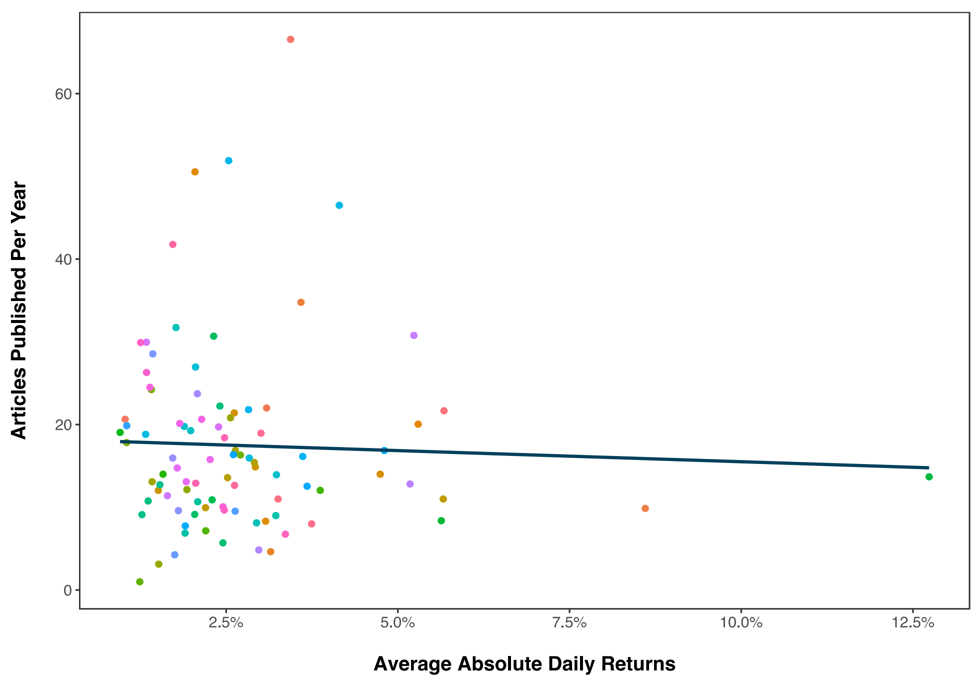

Figure 4: Relationship between Average Absolute Daily Returns and Articles Published per Year

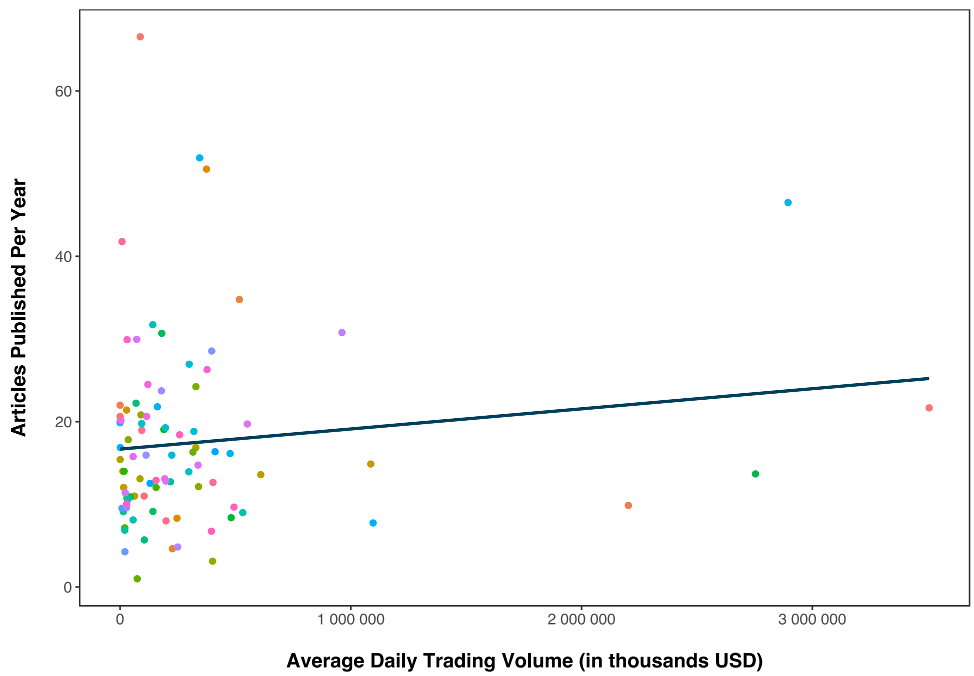

Figure 5: Relationship between Average Daily Volume and Articles Published per Year

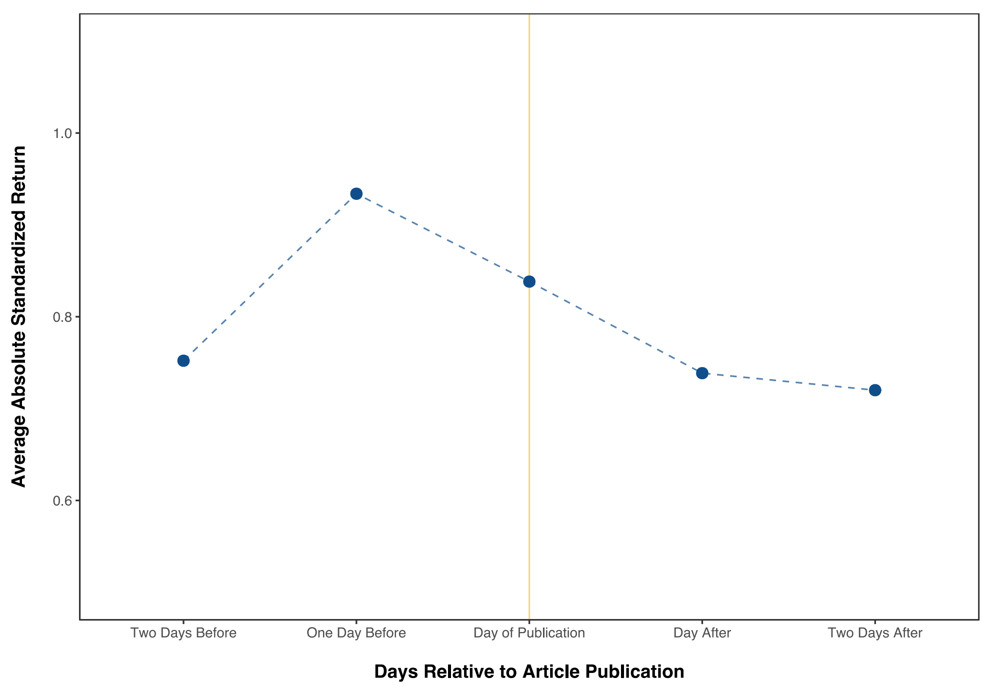

Figure 6: Average Absolute Returns in Days before, on and after an Article Publication

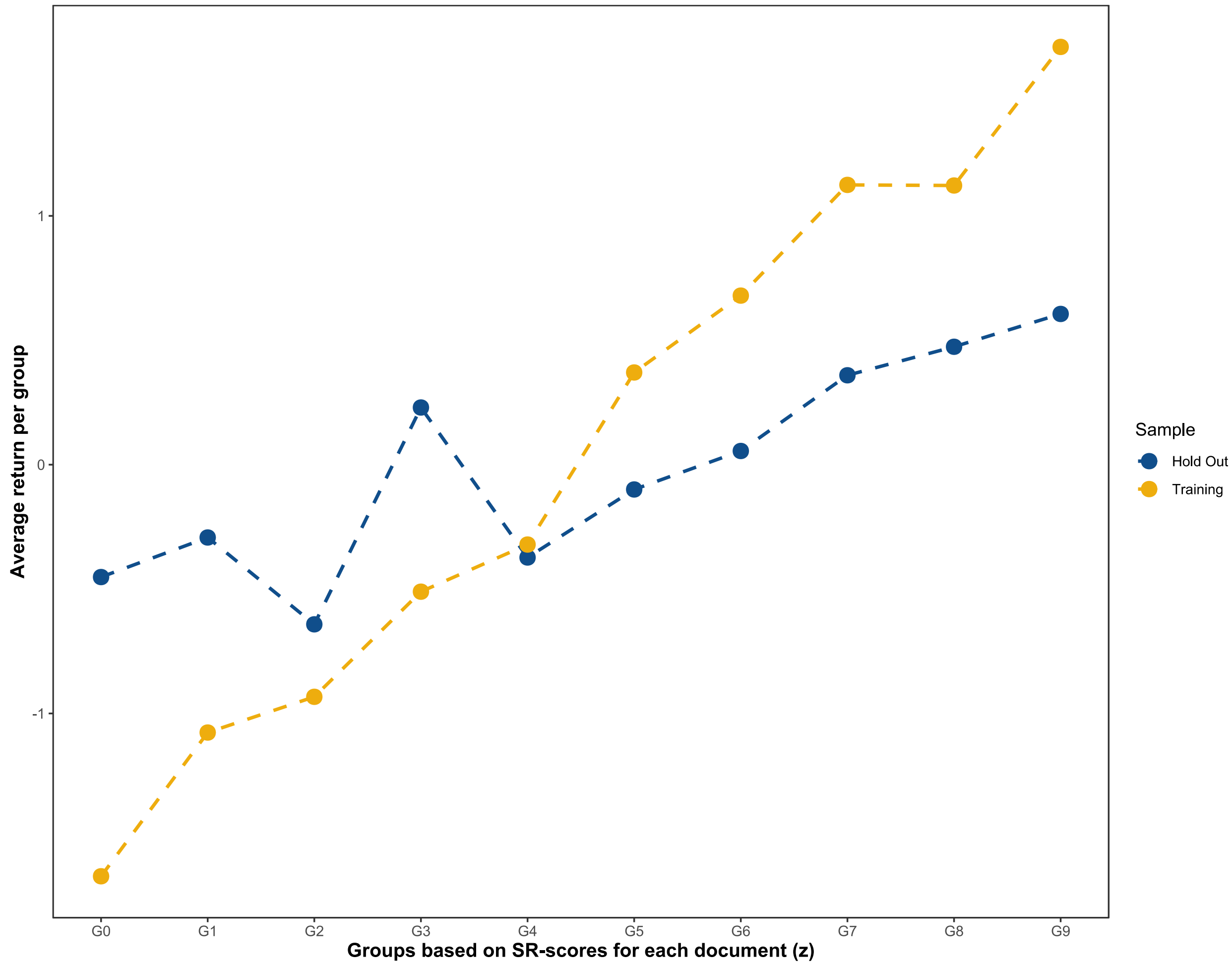

Figure 7: BIGRAM

In this Figure we split the articles by the estimated z_{it} in 10 quantiles – ranging from the 10th to the 90th percentile and compute the average return per group.

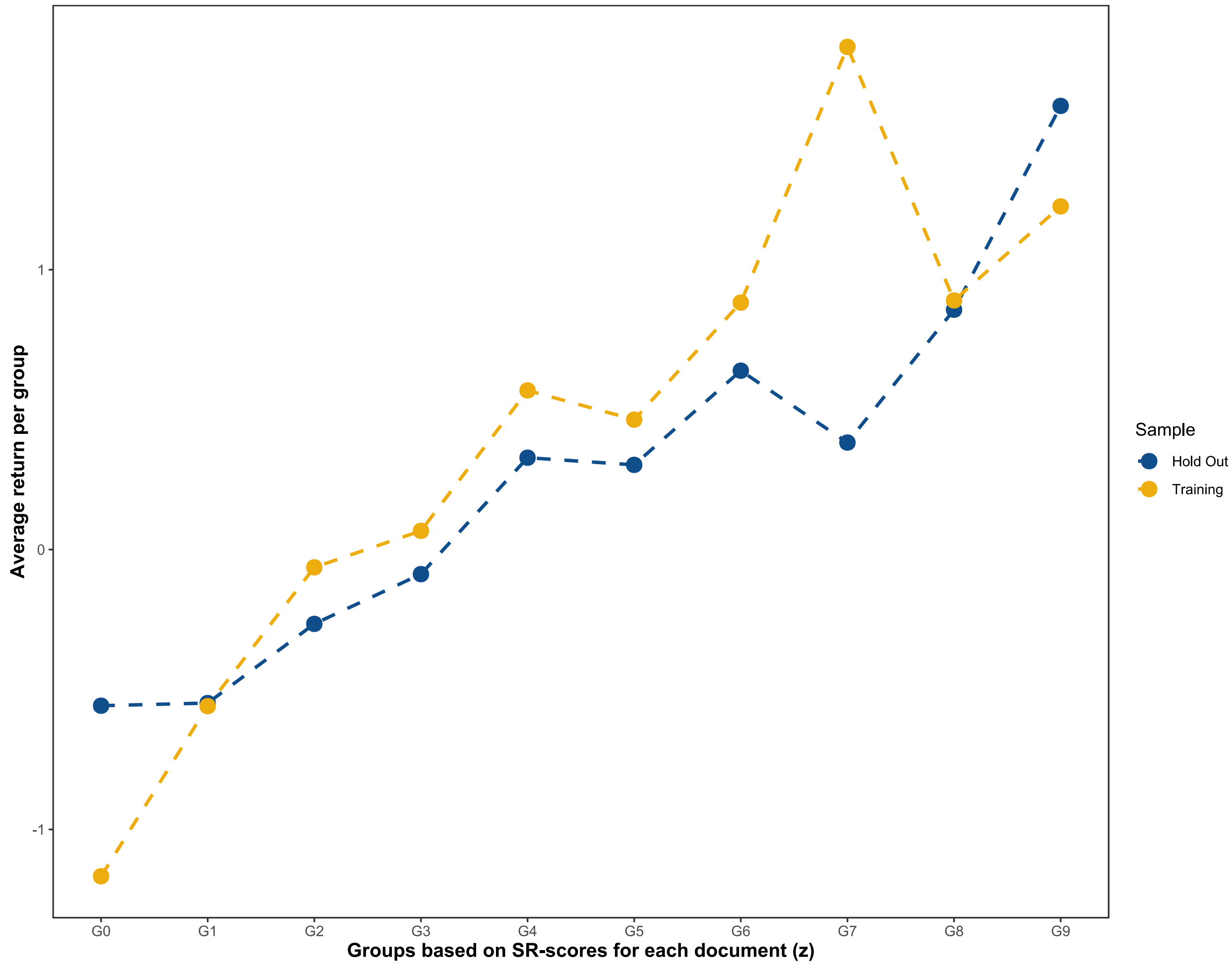

Figure 8: UNIGRAMS

In this Figure we split the articles by the estimated z_{it} in 10 quantiles – ranging from the 10^{th} to the 90^{th} percentile and compute the average return per group.

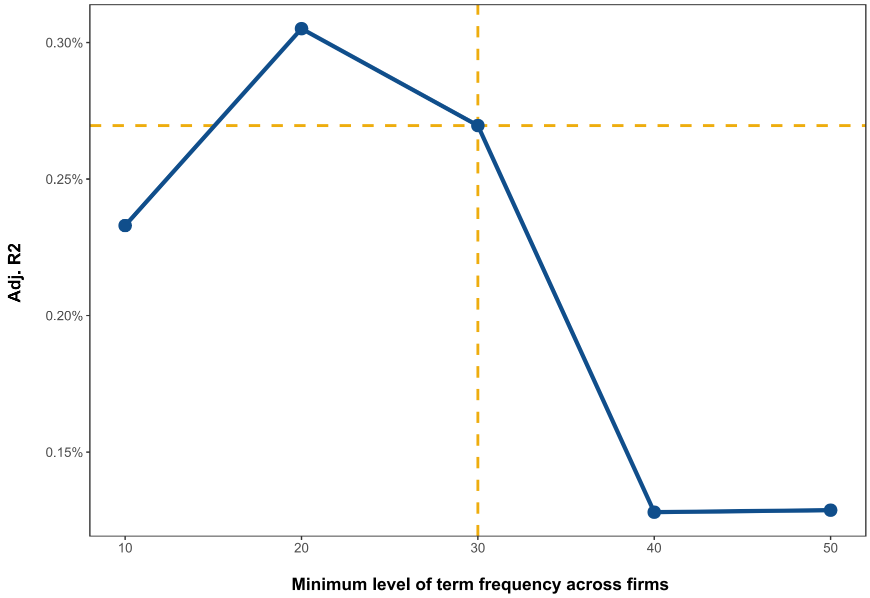

Figure 9: Sensitivity on bigrams – Group B (minimum values of term frequency across firms)

This figure displays how our model performs (using $Adj. R^2$) when changing the parameter minimum values of term frequency across firms – ceteris paribus. Note that the regression used is the same as in Table 9. The parameter (x-axis) removes firm specific effects, meaning that when the parameter is set low, the term only appears in news articles related to few firms. The yellow dashed lines show the parameter level and $adj. R^2$ for our original model. The Figure is made ex-post and shows that when we increase the minimum value of term frequency across firms, the explanatory power increases.

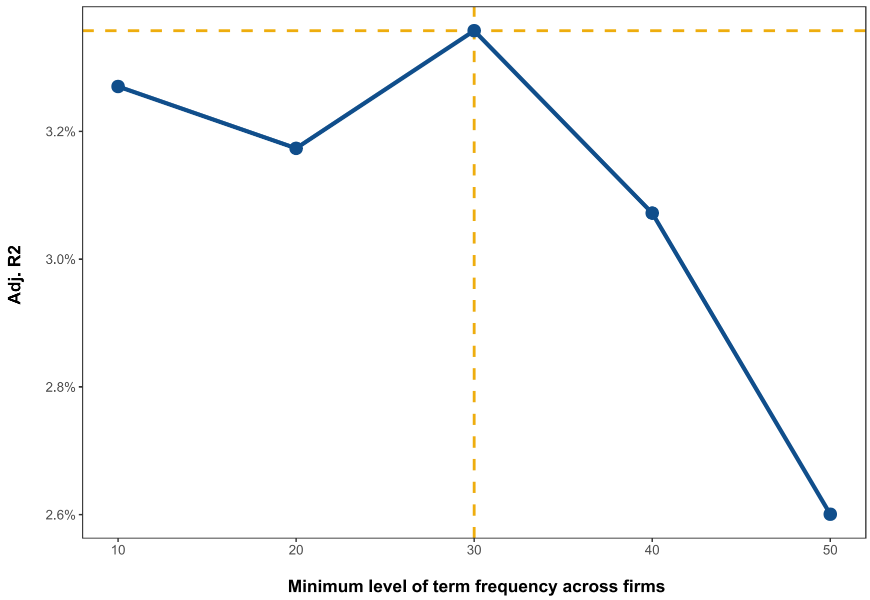

Figure 10: Sensitivity on bigrams – Group A (minimum values of term frequency across firms)

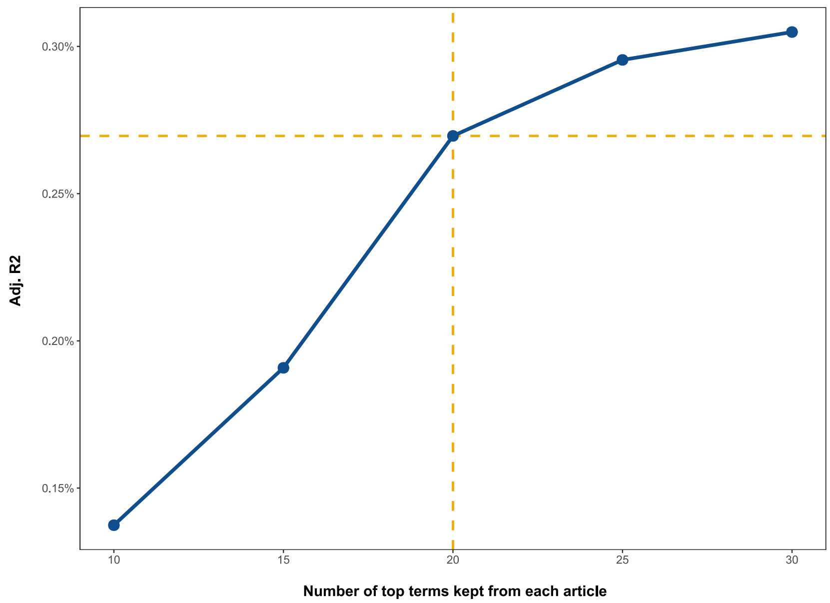

Figure 11: Sensitivity on bigrams – Group B (Number of top terms kept from each article)

This figure displays how our model performs (using $Adj. R^2$) when changing the number of top terms we use from each article – ceteris paribus. Note that the regression used is the same as in Table 9. The yellow dashed lines show the parameter level and $adj. R^2$ for our original model. The Figure shows that the model improves when we increase the top terms from each article. The marginal improvement of increasing top terms flatten out after 20-25 terms.

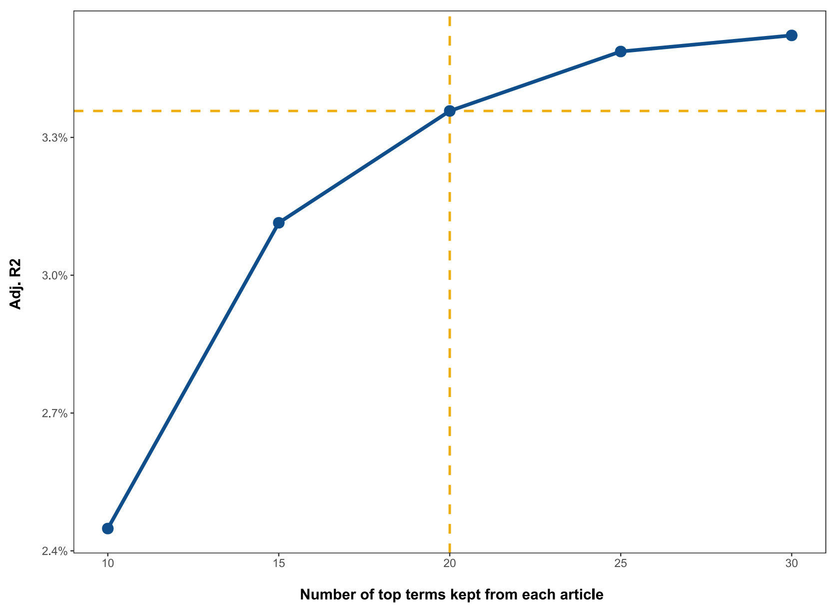

Figure 12: Sensitivity on bigrams – Group A (Number of top terms kept from each article)