BAN432

Applied Textual Data Analysis
for Business and Finance

Final Project
14.11.22 – 21.11.22

NHH

Candidate Numbers:

7, 26, 59, 65

Market Reactions to Wall Street Journal Articles - a Sentiment Analysis

Introduction

In this paper we will make use of sentiment analysis in order to evaluate how the stock market responds to the information provided by articles in newspapers. Information is a valuable asset on a stock exchange which guides investors and their decisions. The goal of the paper is to create a corpus which is able to determine if an article will be received positively or negatively in the stock market. We will base this corpus on Wall Street Journal articles from 2000 until 2022 and use the stock return of the mentioned companies in the same time period to determine whether the market interpreted the news positively or negatively. The finished corpus will be applied to a set of earning call transcripts from 2005 - 2022 in order to assess the external validity of the sentiment analysis.

Task 1

Explorative Analysis

Initial evaluations of the data provided is crucial if they are to be used in further analysis. If any insight is to be gained from the sentiment analysis in this paper it is important that the data used in producing it is explained in a manner which reduces the chance of any biases impacting our conclusions.

Given our comparatively limited set of articles from only one newspaper, it is important to explore which variables take part in determining the amount of news coverage of a given firm. If coverage is skewed toward a certain type of company, it is possible for our corpus to become overfitted to the training set. Plotting the mean articles published for a given firm over the course of their listing on the stock exchange against average volume and return in the same time period can be used to take such predispositions into account in our analysis.

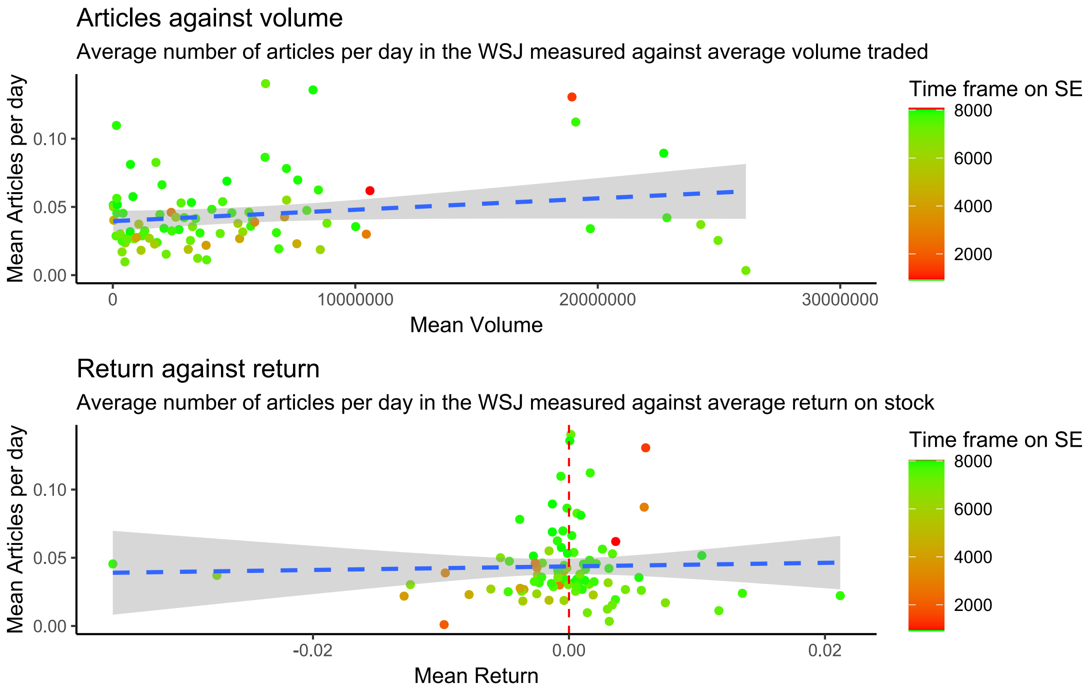
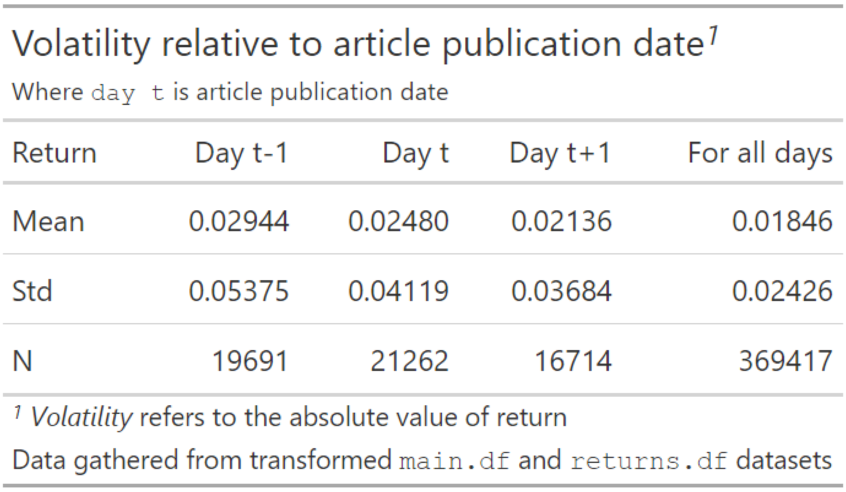


Figure 1 - Articles plotted against volume and return

Figure 1 shows that company stocks with a higher trading volume on average gets covered more than low volume stocks. On the other hand, there are no indicative trends in the data pointing towards companies with very high or low returns being covered at a higher rate than lower volatility companies. A high percentage of the outliers in the data lies on the lower scale of days listed on the stock exchange, which because of the law of small numbers might explain their higher degree of fluctuations.

Given WSJ's role as an independent news provider, covering the stocks which have the largest trading volume would make sense in order to reach the largest possible audience. Lower volume stocks just might not have enough interested parties to warrant a high article turnout. The graph with returns as the explanatory variable shown in Figure 1 points to there being little to no correlation between returns and written articles about a firm on average. This could be helpful in analysing returns based on sentiment later. If WSJ systematically wrote more articles about companies with high or low returns, then neutral words used in a lot of articles would be distorted toward either side of the spectrum. A more even distribution of the data points results in it becoming statistical noise in the sentiment analysis.

Another relevant aspect of the data set is the relationship between a published article and activity in the market. In order to properly model a sentiment analysis, there needs to be a correlation between the articles and market activity. The stock market is often influenced by emotional responses to a given event (Strycharz, Strauss, & Trilling, 2018). The manner in which newspapers present such events might therefore influence the market's emotional responses. Although some articles might be considered a driving factor in significant movements in the market, it would be reasonable to assume that on average there are more independent external factors and events which drive the market change and cause articles to be written.



Volatility relative to article publication date[1]

Where `day t` is article publication date

Return	Day t-1	Day t	Day t+1	For all days
Mean	0.02944	0.02480	0.02136	0.01846
Std	0.05375	0.04119	0.03684	0.02426
N	19691	21262	16714	369417

[1] *Volatility* refers to the absolute value of return

Data gathered from transformed `main.df` and `returns.df` datasets

Table 1 - Volatility around article published

This hypothesis is backed up by the data presented in Table 1, which measures market volatility the day of an article being written, the day before and the day after. There is significantly more volatility in the market the day before an article is written, then on other days. Results from a two-sided t-test tell us that the means for the days around an article being published are significantly different from the mean for all days. This indicated that there are external factors which influence the mean volatility around the time of articles being published. Going forward this analysis will be the basis of how we define market reaction to an article, as the evidence points to the reaction to “news” i.e. the event which the article covers on average takes place the day before the article is published.

Task 2

Method

Constructing a sentiment analysis requires some rudimentary cleaning of the text which is fed to the regression model. Cleaning the text of stopwords, white spaces, numbers and punctuations expands the predictive power of the analysis, as it filters out information which does not contribute to the sentiment of the text. Bounds in the sentiment dictionary are used to prevent inserting a sentiment value that does not occur within a minimum and maximum set of documents. Doing this reduces the overfitting of the training set, as rare words would have been assigned unrepresentative sentiment values based on the training set, and very common words would have noised out the real important words. Alternatively, too strict bounds would filter out words that could help our model. The bounds are therefore set as a trade-off between overfitting and over filtering. Similarly bounds are used to limit the upper and lower length of the words included in the analysis. The minimum word length ensures that the words are meaningful. The maximum word length filters out long and uncommon words that could create statistical noise. After cleaning the text, it gets indexed into a document term matrix which is the required input format in the regression analysis.

The aforementioned regression forms the basis of our further analysis in the paper. Although there are several tools which can be applied to the task at hand, the Multinomial Inverse Regression (MNIR) fits the specific requirements of this analysis. MNIR allows the application of predictive attributes to words in the document term matrix (Taddy, 2013). Estimates of individual words are given independent of each other, and a conjecture is therefore made in regard to the length of the document. The output of the regression is a measure of co-occurrence of a given token with returns and is defined as a MNIR coefficient (Rohrer & Langerfeld, 2022). The MNIR coefficient for a given token indicates a correlation between the term and positive returns in the market. In this specific case it measures how the market in general reacts to a given term used in a Wall Street Journal article and the strength of this correlation. Articles are lagged compared to returns, as Task 1 showed that the greatest connection between returns and articles were returns on days before articles were published. It seems logical that returns should affect the wording of articles to a higher degree than the opposite.

In order to construct a MNIR model and utilize it in order to gain informative insight into the task at hand, certain parameters need to be implemented. Such parameters need to be accounted for in order to specify the exact parameters needed to recreate our results. The word count interval determines the length of the texts to be included in the analysis. An interval is needed for creating a homogenous sample size. This comes with a trade-off, that articles outside do not get accounted for. But on the other hand, the explanatory power of the observations we keep becomes much stronger.

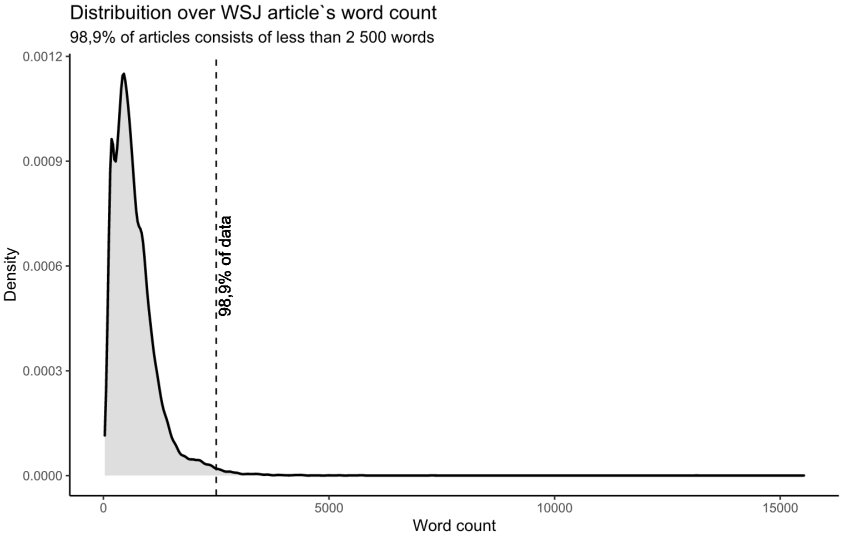
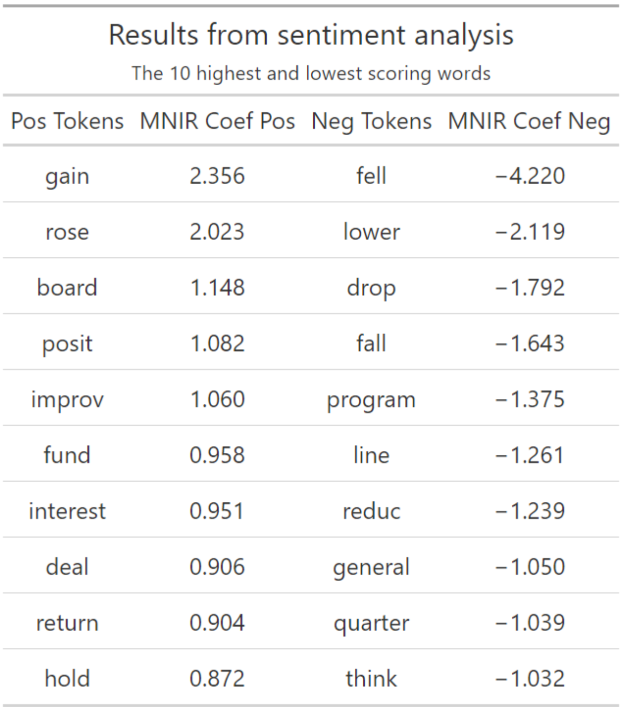


Figure 2 - Distribution of word count

The document term matrix in our model stems the words in the text files. By stemming the words itself they get compressed to their stem. This makes it possible to gather similar words into one stem and connect sentiment on each stem instead of each word. Advantages with this method is that it works well when the connection between returns and text sentiment is quite weak. The model is susceptible to noise when the data set size is limited and having fewer variants of each word helps reduce noise. Different variations of words sometimes have different sentiments. Stemming therefore risks reducing nuances that would otherwise appear. This is less relevant for our model than noise reduction.

N-gram usage has a similar trade-off. Using bigrams instead of unigrams helps distinguish situations where negation is used. The model would e.g. be able to distinguish between 'good' and 'not good' with bi-grams, but not with unigrams. However, using bigrams also creates more variations of word combinations. It is desirable to avoid many variations, as to reduce noise, since there is a limited data set, and a small correlation between word sentiment and return.



Results from sentiment analysis

The 10 highest and lowest scoring words

Pos Tokens	MNIR Coef Pos	Neg Tokens	MNIR Coef Neg
gain	2.356	fell	−4.220
rose	2.023	lower	−2.119
board	1.148	drop	−1.792
posit	1.082	fall	−1.643
improv	1.060	program	−1.375
fund	0.958	line	−1.261
interest	0.951	reduc	−1.239
deal	0.906	general	−1.050
return	0.904	quarter	−1.039
hold	0.872	think	−1.032

Table 2 - Highest coefficients from corpus

Table illustrated below shows the 10 highest and lowest scoring words in the sentiment analysis. At the side, the MNIR coefficient of the given word is presented. Different seeds used in the regression produce similar results, which indicate consistency within the model.

Task 3

Internal Validity

Testing the internal validity of the model requires that the complete data set were split up into a training (Group A) and testing (Group B) set, in which the model is based upon the training set. The relationship between sentiment and return for group A and group B is shown as a regression summary in Table 3 and Figure 3. Group A has approximately twice as high explanatory power and coefficient size as group B, regardless of the seed that is used for the subsets. The imparity indicates overfitting, which means that the model is too largely adapted to the observations in the training set, Group A, and is therefore not equally able to accurately estimate the relationship between return and sentiment in the test set, Group B. Nevertheless, both z-coefficients are significant, meaning that sentiment explains the variation in returns to a small degree in both groups.

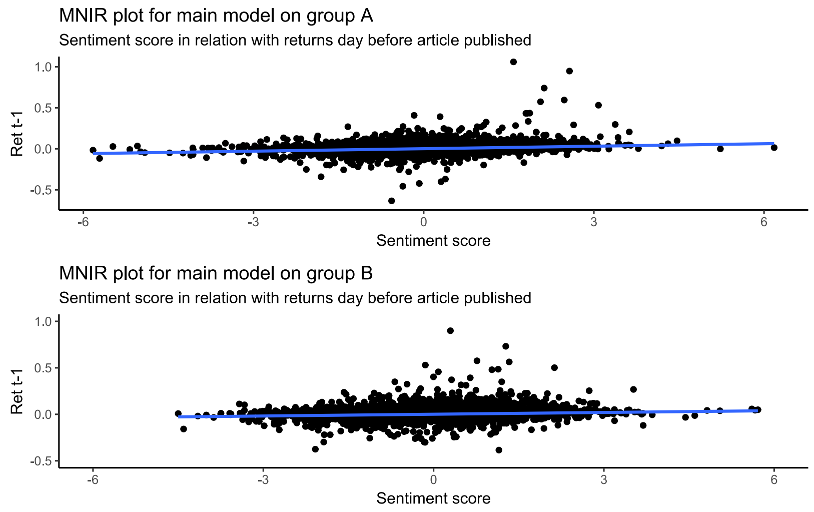


Figure 3 - Sentiment output for Group A and B

	Main model	
	Ret t-1	
	(1)	(2)
Sentiment	0.010***	0.006***
	t = 16.980	t = 11.930
Constant	0.001**	0.001
	t = 2.319	t = 1.530
Test or Train?	Test	Train
Observations	6,834	8,212
R^2	0.040	0.017
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3 - Regression output

Table 3 shows Group B split into two subgroups with companies that have few and many news articles respectively and regression run on both groups. The model appears to be somewhat inconsistent regarding coefficient size and explanatory power between the two groups. This was discovered by running the model with different seeds. They are however always in a similar range, and significant. The model could be systematically better at finding the connection for e.g. companies with many articles. This does not appear clearly, however, as the statistical values vary from seed to seed. Overall, it is reasonably good at finding the connection between return and sentiment for both groups, with some variation with each seed. The internal validity therefore seems acceptable.

	Regression splitted on n articles	
	Ret t-1	
	Few articles	Many articles
	(1)	(2)
Sentiment	0.007***	0.006***
	t = 5.315	t = 10.930
Constant	0.002	0.0005
	t = 1.332	t = 0.848
Test or train?	Test	Test
Observations	2,058	6,154
R^2	0.014	0.019
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 4 - Regression output

Performance Test 1: Changing Bounds

Bounds define the interval for how many texts a word can appear in, and still be added to the document term matrix. The original bounds are between 2500 and 5500 texts, out of the around 15 000 texts in the data frame. Words that appear in relatively few or many texts are filtered out. This test will try to expand the bounds in both directions, so that it includes more words in the document term matrix. Bounds are expanded to between 50 and 8000, which means that the document term matrix now includes all words except the most rare and common ones. By running the model again with these changes, the explanatory power for Group A, i.e., the in-sample set, has increased greatly. Conversely, the explanatory power of the model for Group B has fallen to zero. This indicates overfitting, meaning that the new words in the document term matrix appear to correlate with return in Group A, without having the same relationship in Group B. The model's prediction capabilities are therefore weakened by the increased bounds.

Performance Test 2: Word Length

The word length interval parameter will also be changed to test the performance of the model. It is initially between four and twenty characters, but in the test, it will be reduced to between five and ten characters. The effect of this restriction is that fewer words qualify to be included in the document term matrix due to their character length. This is regardless of whether they could otherwise have been relevant. The result of the regression shows that the analyses of both Group A and B have been weakened. The explanatory power for both is greatly reduced, and for the test set it has become approximately zero. The sentiment variable is also no longer significant in the test group. Overall, further restricting the interval parameter for word length seems to greatly reduce the model's performance.

Performance Test 3: Word Count

The word count interval is originally between 240 and 2500 words, which means that texts that are longer or shorter than this are filtered out. Since the outliers are filtered out, the text set used in the model is more homogenous, which is advantageous. This could lead to the model finding more correlation in the texts and return data that is kept. In this test, the upper and lower limit for word count is removed, which means that the texts will be included, regardless of length. By running the regression again, it appears that the explanatory power has fallen in both group A and group B. This is probably because long and short texts do not have the same relationship between return and sentiment, as the text with average length. On the other hand, both groups' sentiment coefficients are still significant, so the connection is still present. The result is not surprising, since there are now included more statistical outliers in the data set, which makes it harder to find correlation. Overall, it is difficult to decide whether changing the word count parameter makes the model worse. On one hand the explanatory power has dropped sharply on the test set. However, it also includes a larger share of the texts provided in the task and could thus be more versatile.

	Regressions with parameter tuning							
	Results from tuned parameters juxtaposed against the main model							
	Main Model		Bounds		Word Length		Word Count	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sentiment	0.010***	0.006***	0.023***	0.0004	0.004***	-0.0003	0.009***	0.003***
	[16.979]	[11.927]	[46.192]	[0.646]	[7.113]	[-0.578]	[14.252]	[7.448]
Constant	0.001**	0.001	0.0004	0.001*	0.001**	0.001*	0.002***	0.001*
	[2.319]	[1.530]	[0.683]	[1.736]	[2.048]	[1.724]	[3.081]	[1.781]
Test or Train?	Test	Train	Train	Test	Train	Test	Train	Test
Observations	6,834	8,212	6,834	8,212	6,834	8,212	8,013	9,818
R^2	0.040	0.017	0.238	0.0001	0.007	0.00004	0.025	0.006
Note:	*p<0.1; **p<0.05; ***p<0.01							

Table 5 - Regression output for performance tests

Task 4

External Validity

To measure sentiment in earnings calls and determine a correlation to stock market returns similar pre-processing to constructing the sentiment dictionary is conducted. The transcripts of earning calls consist of corpuses from both introductory and question and answers (Q&A), and they are cleaned with the same procedure as the WSJ articles.

Each corpus and transcript are indexed into a document term matrix which is used to compute the relative usage of terms in the constructed sentiment dictionary. The sentiment dictionary consists of terms and their sentiment weight measured by MNIR coefficients. For each document the relative usage of each term that exists both in the document term matrix and in the sentiment, dictionary is computed. The relative usage is measured against all other terms in the document term matrix used in a transcript. This is computed once for positive and negative terms which results in each transcript introduction and Q&A for each date receiving a positive and negative sentiment score.

Linear regression is fitted to each transcript by using the sentiment score as the explanatory variable and the return registered on the same date as the output variable in order to measure correlation between the two variables.

Model Tuning

There are fewer articles in the earning calls set than in the main set, which means that the bounds must be adjusted. The bounds are tuned down significantly, to a reasonable interval. The sentiment value of both positive and negative words from the Q&A is included in the regression because it is interesting to see if the wording of the discussion is correlated with returns. The introduction text sentiment variables are excluded from the model, as the texts appear relatively standardized and uninformative regarding the state of affairs and returns of the company. Returns are no longer lagged in relation to the texts because the earning calls should come before the changes in the share price.

Results

The results show that the positive Q&A variable is significant, but the negative variable is not. After further investigation, it turns out that the word "boost", which has strong positive sentiment regardless of seed, appears in the earnings calls and provides a significant connection. However, there are no strong negative words that appear in the set; sometimes some negative words with weak sentiment appear, but these are not consistent across different seeds, suggesting that they could be noise. It is therefore not possible to consistently find a connection between negative sentiment and returns in the earnings calls set. The fact that strongly negative words do not appear in the earning calls, may suggest that managers deliberately avoid using them, but this cannot be concluded with certainty. The explanatory power of the model is slightly higher than in the main model, which may indicate that the connection between the choice of words in earnings calls and return is stronger than the connection with news articles.

The earning calls model was split between the companies in Group A and Group B. The results show that positive sentiment for Q&A is only significant for the companies in the test set in the main model, which does not seem logical. This is probably since the texts with the strongly positive word "boost" only appear in one group, which in return becomes significant.

Compared to the results in task 3, the model performs significantly worse. Where the model in Table 3 had significant coefficient sentiment variables and high explanatory power in both groups, the model in Table 6 only has significant coefficients for the positive Q&A sentiment in one of the groups. The external validity is therefore significantly weaker than the internal, which means that the model is much better at predicting relationships internally in the main data set than for the earnings calls. There could be several reasons for this. First, the earning calls set size is much smaller than the main set, which can make it difficult to predict with it. Furthermore, very few words with sentiment value appear in the earning calls set. In addition, there may be systematic differences in how CEOs and CFOs word themselves in earning calls compared to the wording of independent Wall Street Journal articles. These factors would indicate lower external validity, even though the choice of words in the earnings calls may actually have a greater connection with the share price than the articles.

	Transcript of earning calls and stock return		
	Ret t		
	Full sample	Splitted sample	
	(1)	(2)	(3)
Negative sentiment	0.136	0.832	-0.044
	[0.204]	[0.568]	[-0.058]
Postive sentiment	1.631**	1.040	2.016**
	[2.431]	[0.943]	[2.358]
Constant	-0.0005	0.001	-0.002
	[-0.118]	[0.217]	[-0.388]
Test or train?	Full set	Training	Test
Observations	319	157	162
R^2	0.018	0.007	0.034
Adjusted R^2	0.012	-0.006	0.022
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 6 - Regression output

Conclusion

The goal of this paper was to construct a corpus dictionary targeted at newspaper articles. In summary, our corpus was able to predict some of the variation in the stock market based on the words used in an article. However, the choice of words captures only a small part of the variation in stock prices. The low explanatory power is probably due to the stock prices' extreme volatility, and the relative limitations of the data set. The model manages to find correlation in both the training set and the test set, although it is stronger on the training set. This indicates some overfitting in the model. The internal validity of the model seems acceptable since the division between companies with short and long texts in the test set still gives significant correlation and similar explanatory powers in the two groups. The link between sentiment in earning calls and return is captured less effectively by the model. A certain relationship does emerge, but not for all variables. The external validity for financial sentiment seems weaker outside of Wall Street Journal articles, which points to the corpus being sensitive to the dataset it is applied to.

References

Rohrer, M., & Langerfeld, C. (2022). Lecture 15: Document Clustering - MNIR. *[PowerPoint and R-script]*, Accessed through Canvas.

Strycharz, J., Strauss, N., & Trilling, D. (2018). The Role of Media Coverage in Explaining Stock Market Fluctuations: Insights for Strategic Financial Communication. *International Journal of Strategic Communication*, 67-85.

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, 755-770.