Midterm Project Report

Summary Section

The goal of this project was to analyze the 10k reports of all of the companies in the s&p500 in 2022 to determine whether the sentiment of the text had an impact on their stock price in the days following. The specific question my project addressed was whether the mentioning of divestitures, lawsuits, or the release of a new product, surrounded by negative or positive sentiment, had correlation with the change in stock price. For divestitures, I used the following phrases to search the documents for a sell off: Divestiture, Disposal, Liquidation, Sale, Offloading, Exit, De-merger. For lawsuits, I used the following words: Litigation, Lawsuit, Arbitration, Settlement, Regulatory, Investigation, Class Action, Subpoena, Injunction. Lastly, for the relase of a new product I used the following phrases: Product Launch, Rollout, Market Introduction, Debut, Commercialization, Unveiling, Deployment, Soft Launch, New Product. The results came back that the more positive and negative sentiment words did not have an impact the returns of the company. However, the mentioning of the divestiture and lawsuits did have correlation. The release of a new product had the most impact to stock price as this caused the stock to decrease.

Data Section

To get the returns of the comapany, I pulled the closing stock price of the company on the day of their release and compared it to the stock price of the company 4 days after the 10k released. Firstly, to obtain the date of the companies 10k reslease, I used the accession number, a value that is given to a documents release that identifies its time of release and the company as well. The process in doing this began with pulling all of the companies fro the s&p 500 from wikipedia and downloading the csv file. I then used the CIK, an identification code for each company, from the wikipedia page as well as a start date of 1/1/2022 and end date of 12/31/2022 to identify the 10k in the 2022 year. This provided the accession number and allowed for the download of each companies 10k. The code below used the accession number to identify the date of release of the 10k:

```
dates = r.html.find('#contentDiv > div:nth-child(1) > div.formContent > div:nth-child(1) > div:nth-
child(2)',first=True).text
```

The next step was to identify the date of 4 days after the release of the 10k. This was difficult because 4 days after means 4 trading days after, so I needed to account for weekends and holidays. To do this, I found a helpful stackoverflow that used "import pandas_market_calendars as mcal". The code below is what was used to find the date 4 trading days after:

```
nyse = mcal.get_calendar('NYSE')

def get_next_4_trading_days(filing_date):
    """
    Given a filing_date (str or Timestamp), returns the 4th
    trading day after that date on the NYSE calendar.
    """
    filing_date = pd.to_datetime(filing_date).tz_localize('UTC')
    start_date   = filing_date
    end_date     = filing_date + pd.Timedelta(days=10)

    valid_days       = nyse.valid_days(start_date=start_date, end_date=end_date)
    days_after_filing = valid_days[ valid_days > filing_date ]
    return days_after_filing[3]
```

To briefly breakdown the code, the nyse gets the calendar year for all trading days in a given year. The def function inputs the filing date, which is the date of the release of each respective 10k, and after some data changing, inputs it as the start date. The end date is 10 days after to overestimate time like 3 day weekends, or anything along those lines. Then, the valid_trading_days finds the valid trading days within those 10 days, which then the followed code finds the date of 4 days after. The reason 4 trading days was chosen was because one day after generally has a reactionary spike to certain things in the 10k or earnings call, while the following 2nd and 3rd day allow for correction. The price of the stock on the fourth day has the highest representation of th reaction to the sentiment of the 10k, rather then just the primary glance of it and the financials. Any time following that inputs the issue of external factors that have no correlation with the 10k, like a factory burning down, or some extremity not associated to the sentiment of the 10k. To obtain the sentiment variables, the code below was used for the LM and the BHR, only positive is shown, but negative was used as well:

```
file_path = "inputs/LM_MasterDictionary_1993-2021.csv"

df = pd.read_csv(file_path)

LM_positive = df[df['Positive']>0]['Word'].tolist()

LM_positive = [e.lower() for e in LM_positive]

LM_positive_pattern = r'\b(?:' + '|'.join(map(re.escape, LM_positive)) + r')\b'

_____
_____

with open('inputs/ML_positive_unigram.txt', 'r') as file:

    BHR_positive = [line.strip().lower() for line in file]
```

```
BHR_positive_pattern = r'\b(?:' + '|'.join(map(re.escape, BHR_positive)) + r')\b'
```

For the LM code, I filed a path to obtain the dictionary, then identified the positive column and added each of the words in the column to a list. This list was then made all lower case to account for case sensitivity issues in the matching process. For BHR, the process was a bit simplier as it was aleady in a .txt file, so I simply stripped each word, made them lowercase, and added it to a list. Both the LM and the BHR have a pattern line which is needed for using the NEAR function later in the code, essentially a way for the code to use the words individually and prevent words that are similar to be used. The number of words in the LM positive dictionary are 347, the LM negative dictionary are 2345, the ML positive dictionary are 75, and the ML negative dictionary are 94. To find contextual sentiment, I used the following funtion: NEAR_finder(newprod_words,BHR_positive,text). This specific code was an example of the new product conetextual sentiment, as newprod_words is the list of words mentioned. I used BHR_positive because based on the comparison article between the ML and the LM lists, the ML llist seemed to prove more recent and accurate to the project. Lastly, the text is the final input, simply the cleaned up text of the respective 10k. To reiterate, the three topics I chose were divestitures, lawsuits, and the release of a new product. Divestures intruiged me because there is positive and negative reactions to the sell off of a subsidiary, depending on the deal, so identifying the sentiment around it could account for the stock change after the release of the 10k. With lawsuits, these could be lawsuits the company undergos, or lawsuits they impose on others. The difference between the two are identified through the sentiment around it, so seeing how th stock reacts would be in interesting thing to identify. Thirdly, the release of a new product can be viewed as positive or negative, whether or not investors believe the product will do well affects the stock price. So, to determimine the psychology of investors with a new product, there may be an impact with the sentiment used near the products release. Below are the summary statistics for my data:

Statistic	CIK	Returns	positiveLM_words	negativeLM_words	positiveBHR_words	negativeBHR_words	positiveDivestiture
count	500.000000	483.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	790452.900000	0.002213	0.004987	0.015916	0.023949	0.025901	0.000117
std	553237.400000	0.046728	0.001316	0.003688	0.003492	0.003391	0.000097
min	1800.000000	−0.358915	0.001226	0.006609	0.007966	0.008953	0.000000
25%	97409.500000	−0.024442	0.004095	0.013297	0.021954	0.023966	0.000054
50%	884064.000000	−0.001899	0.004899	0.015664	0.024122	0.025900	0.000093
75%	1137778.000000	0.025888	0.005662	0.017860	0.026131	0.027813	0.000160
max	1868275.000000	0.202210	0.010899	0.030185	0.037982	0.038030	0.001009

To summarize the importance of this description, the mean of each columns is the following:

Variable	Mean (%)
Returns	0.2213
LM Sentiment	
• positiveLM_words	0.4987
• negativeLM_words	1.5916
BHR Sentiment	
• positiveBHR_words	2.3949
• negativeBHR_words	2.5901
Divestiture Context	
• positiveDivestiture	0.0117
• negativeDivestiture	0.0202
Lawsuit Context	
• positiveLawsuit	0.0184
• negativeLawsuit	0.0421
New-Product Context	
• positiveNewprod	0.0016
• negativeNewprod	0.0024

The return is the average return 4 days after the release of the 10k. The values positiveLM, negativeLM, positiveBHR, and negativeBHR represent the average percent of words in each 10k that are in the respective list. The values for the contextual sentiment variables are the average percent of words in the list of words provided that also have a positive/negative senetiment following it. This passes an initial smell test as all of the values are different and the small percent values make sense as few portions of the text will have the words provideed.

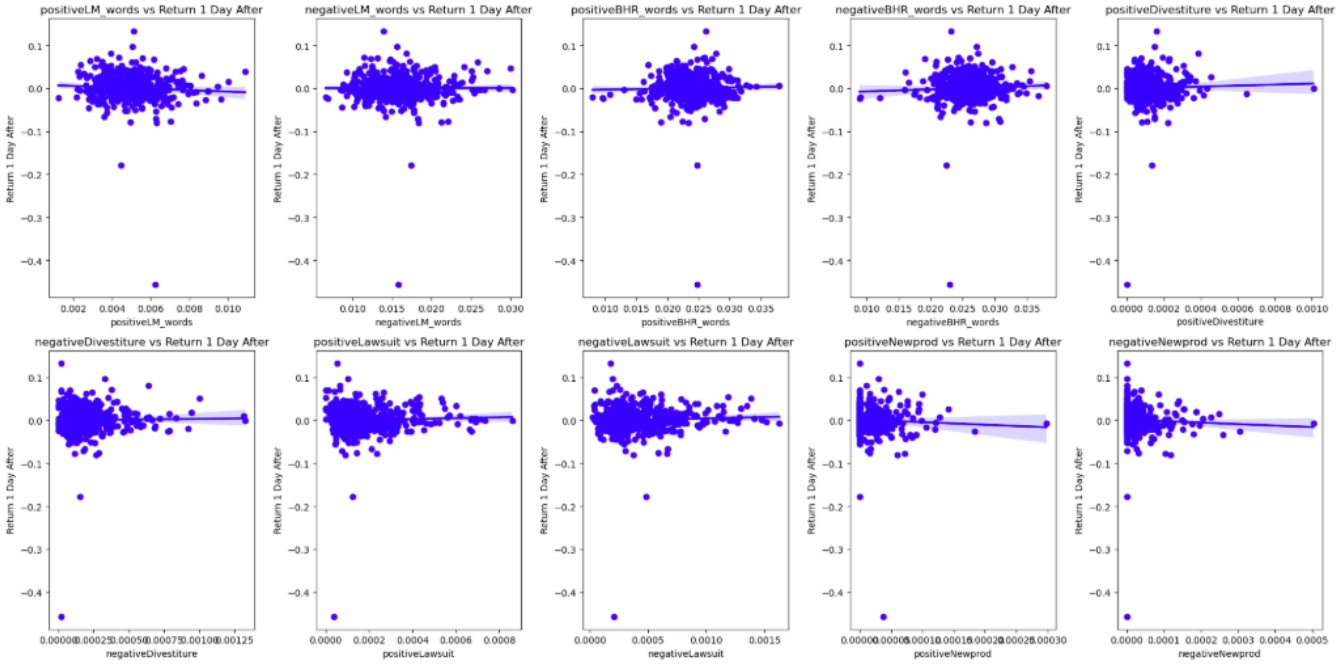
Results

Below is a correlation table with the returns of the firms 4 days after and 1 day after:

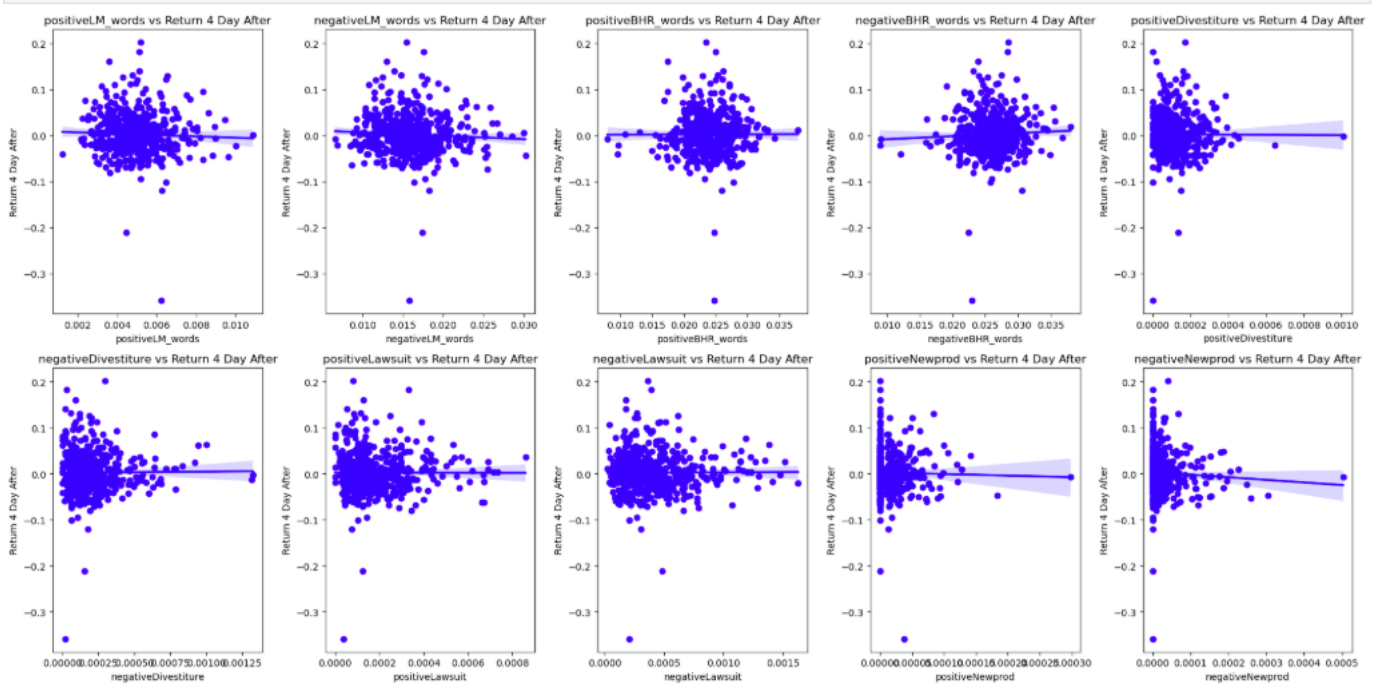
Sentiment Variable	Return 4 Days After	Return 1 Day After

Sentiment Variable	Return 4 Days After	Return 1 Day After
positiveLM_words	-0.040962	-0.063548
negativeLM_words	-0.058524	0.002586
positiveBHR_words	0.002521	0.026742
negativeBHR_words	0.048447	0.049704
positiveDivestiture	-0.003838	0.034160
negativeDivestiture	0.010889	0.015924
positiveLawsuit	-0.001302	0.040303
negativeLawsuit	0.006535	0.049033
positiveNewprod	-0.020575	-0.049029
negativeNewprod	-0.057308	-0.048637

Below is a scatterplot of all of the sentiment variables with 1 day after returns:



Below is a scatterplot of all of the sentiment variables with 4 day after returns:



Discussion Topic 1

The positive LM and negative LM variables had a -.063548 and .002586 correlation with the one day return, respectively. The The positive BHR and negative BHR variables had a .026742 and .049704 correlation with the one day return, respectively. In theory, the positive sentiments should have positve correlation, and the negative sentiments should have negative correlation, meaning positve words make the stock goes up and negative words make the stock go down. This result leads me to believe there isn't much correlation with sentiment and the return of stock prices the trading day after. The difference between the LM and the BHR is quite significant. The LM positve had a slope of -1.64, The LM negative had a slope of .02, The BHR positve had a slope of .26, The BHR negative had a slope of .5. These values analyzed how the increase of words in each respective list would impact returns, so the most significant was the more LM positive words the more negative the stock return was. Again, there isn't much similarity between the LM and the BHR variables.

Discussion Topic 2

My results were different from the ML_JFE.pdf article, as their analysis concluded there was correlation. One possible reason is the text they used. I only used text from 10ks while they used text from earnings call transcripts, 10-K filings, and WSJ articles. This gives them more diversity and possible word counts. They also cleaned up their data more effectively than I did, as they were able to handle outliers, utilize tokenization, stemming, and weighting sentiment scores.

Discussion Topic 3

For divestitures, the positive words had a 0.034160 correlation and an 11.9x slope while the negative words had a 0.015924 and a 3.13x slope. These correlations are a bit low, but I didn't expect high values as the entirety of the stock shift wouldn't be due to simply divestitures. Since the slopes are significant, meaning an increase in positive or negative divestiture words leading to larger returns, there is means to investigate further.

For lawsuits, the positive words had a 0.040303 correlation and an 10.04x slope while the negative words had a 0.049033 and a 6.15x slope. The same logic applies from divestures as there is low correlation, as expected, but the slopes lead me to believe a further look would be beneficial.

For a new product release, the positive words had a -0.049029 correlation and an -60.23x slope while the negative words had a -0.048637 and a -34.88x slope. Both of these correlations are negative and have very large magnitude slopes meaning there is likely a strong influence from the release of a new product. It seems as though regardless of sentiment, there is a negative stock reaction to the release of a new product.

Discussion Topic 4

When it comes to the contextual sentiment, the return correlation between 1 day after and 4 days after is relatively similar. There are only minor changes in the correlation values, while the slopes are somewhat similar as well. For the most part, the correlation values increased in magnitude as time increased, likely meaning the intital spikes were recorrected over time and investors agreed upon the correct stock prices.