Using News Sentiments, Social Media Opinions & Other Companies' Data to Drive Stock Investment Decisions

Mojeed Abisiga

December 2022

Abstract

The world of investment in companies' stocks has always been data-driven, accurate information and analysis underpins every decision made about market fluctuations. In fact, the very nature of this analysis is always changing as new data sources, practices, and insights become available. It is a known fact that every data source can offer investors a competitive edge. For example, analyzing foot data can quickly identify increase or decrease in demand throughout the day; this can help investors understand when these kinds of business are either hitting or missing sales goals.

In view of these, it will be of very big value if investors or potential investors in companies' equity or stocks can visit a one-stop shop to get relevant data, information or insights that can be used in making decisions on the kind of investments they should do or specific companies to focus their investment on. The kind of information that will be of great value to investors include industry and sector of the companies, Company Details (company locations, employee counts, top officers within the company, noteworthy customers etc.), Company Financials (stock prices, market cap, key operating data from financial statements like cash flows, long term debt, capital expenditure, operating expenses, revenue, etc.), Company Sentiments (companieskey news and news sentiment, social media sentiments and topic modelling), other key visualizations & comparative analytics (revenue by location, sector breakdowns, side-by-side comparison between sectors based on KPIs, comparing industries).

Methods

Primary Sources of Data:

The focus for this study was Fortune 500 companies, primarily organizations that are key players in 3 major sectors - Energy, Financial Services, and Technology.Our use case was a deep dive on thirty (30) organizations, spread across these 3 different sectors, with 10 organizations considered per sector.

This case study utilized 2 main data sources which are Stock Analysis, and Twitter. For the Stock Analysis website, the first step that was taken is identifying the required KPIs to solve this problem for stock investors, after which some research was done to find relevant data sources for these KPIs. Also, appropriate measures were taken in verifying from Stock Analysis website's for the terms of Use of the page, to ensure that the data was available for commercial use and if scraping is allowed. The organizations that were considered are:

Financial Services:

- JPMorgan Chase & Co.
- Mastercard
- HSBC Holdings PLC
- Lloyds Banking Group
- Barclays
- American International Group
- Canadian Imperial Bank of Commerce
- American Express
- The Goldman Sachs Group
- Bank of America

Technology:

- VMware
- Atlassian
- Amazon
- Apple
- Broadcom Inc.
- Adobe
- Nokia
- Dell
- Autodesk
- Shopify

Financial Services:

- Duke Energy Corporation
- DTE Energy Company
- Shell
- Exxon Mobil
- BP
- Hess Corporation
- Marathon Oil Corporation
- Ecopetrol S.A.
- PetroChina Co. Ltd.
- Schlumberger

Data Points Scrapped or Generated For Analysis:

Company Details: Name of Company, number of Employees, industry, sector, year company was founded, name of CEO.

Company Financials: Shares Outstanding, revenue, net income, stock prices, capital expenditure, market capitalization, long-term debt, and common stock value

Company Sentiment: Analyst sentiments, noteworthy news updates, news sentiments.

Social Media: Tweets, number of likes, number of retweets, date, location, author of tweet, number of followers, popularity of tweet, relevance of tweet.

Data extraction from Stock Analysis was achieved with R while that from Twitter was achieved with Python. Collected only tweets that span a period of past 2 weeks from when the studies were conducted because of Twitter API free access limitation. In addition, the stock data is a highly volatile data; the market deals with a high level of fluctuation, so it was only reasonable to use news or tweets that were very recent for

our studies. The Twitter handle of these companies were used to extract the recent tweets made by twitter users in response to these companies or talking the companies, for companies who didn't have official handles the most unique word or phrase that descibe the company was used to query the Twitter for tweets about these companies. The extracted data was afterwards, transformed into a more usable format. The scraping from Stock Analysis was done using rvest library after getting the xpath selectors for each of the data points needed right.

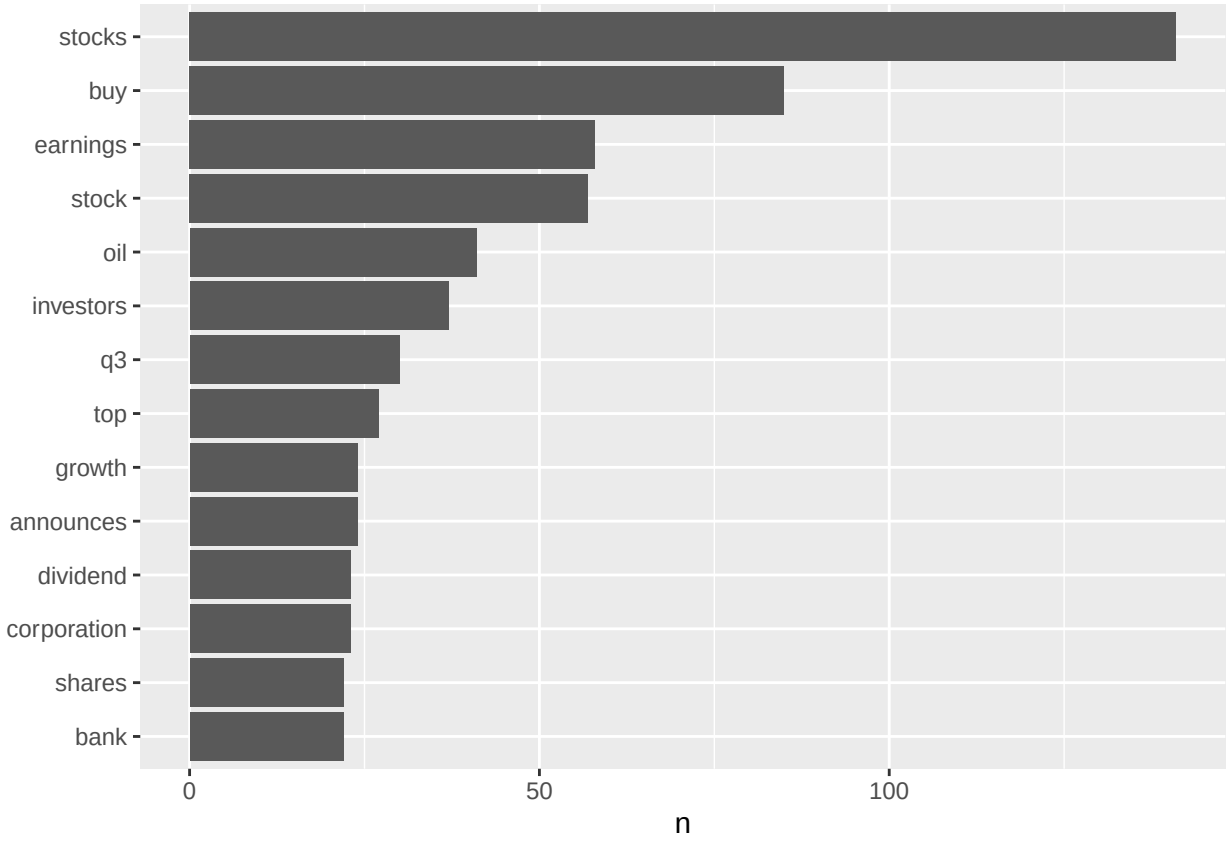
A number of Natural Language Processing (NLP) techniques were used in this study, some of which were Sentiment Analysis, Topic Modelling, N-grams, TF-IDF, and Text Data Wrangling.

Some of the libraries used in R include tidyverse, tidytext, ggplot2, tidyr, forcats, textdata, scales, wordcloud, stringr, tidymodels, httr, purrr, rvest, quanteda, ggraph, igraph, among many others. While some of the libraries used in Python include tweepy, vaderSentiment, unicode, urllib3, pandas, datetime, and numpy

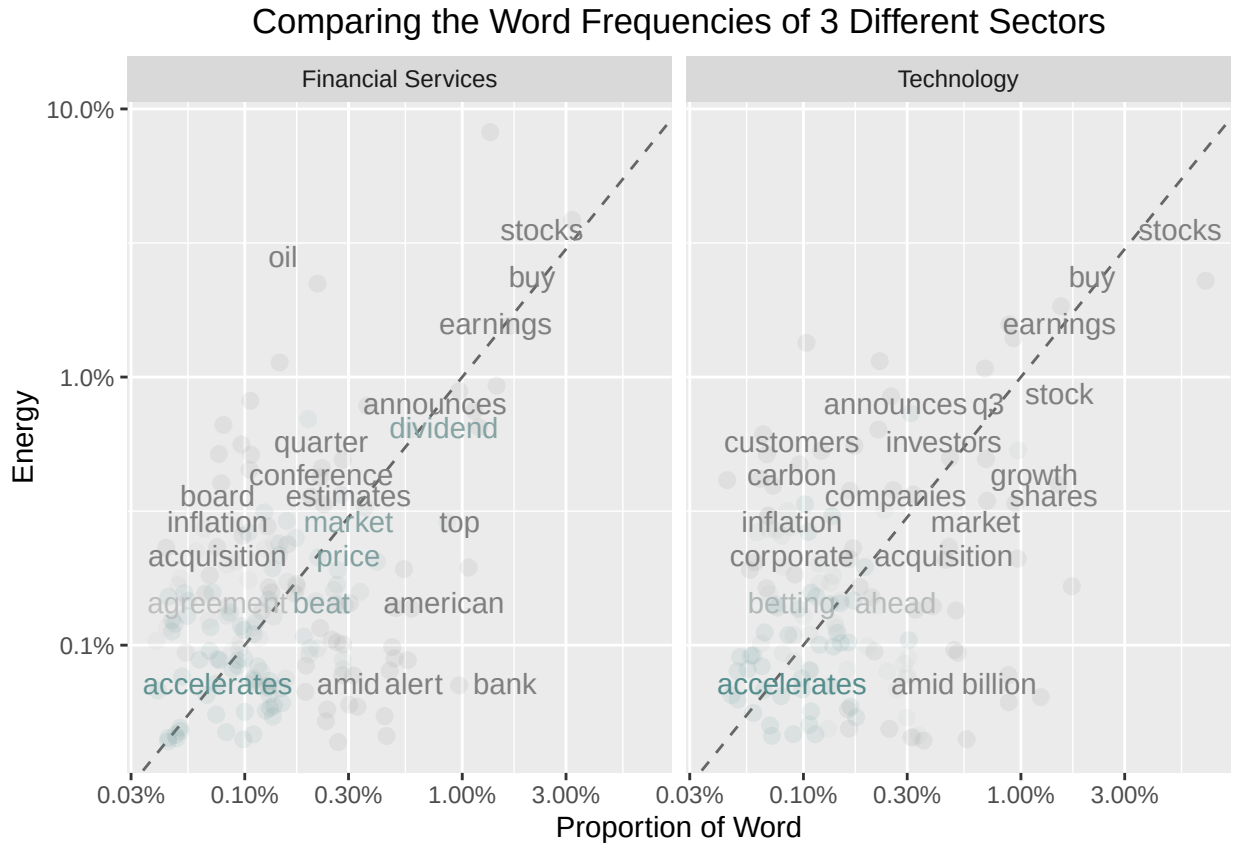
The tweets gotten from Python were exported in CSV file format and then ingested into R where all other analysis and visualizations were done.

Results

Most Frequent words in the News



From the figure above, it can be seen that all news about the companies used in the case study were mostly around stocks, earnings, shares, investors, dividend, which is very much expected since we got the news from stock analysis website. The **q3** indicates that news are kind of centered around the current quarter (third quarter) of the financial year of most companies. The oil companies must be really talked about for it to appear at the top of the chart.



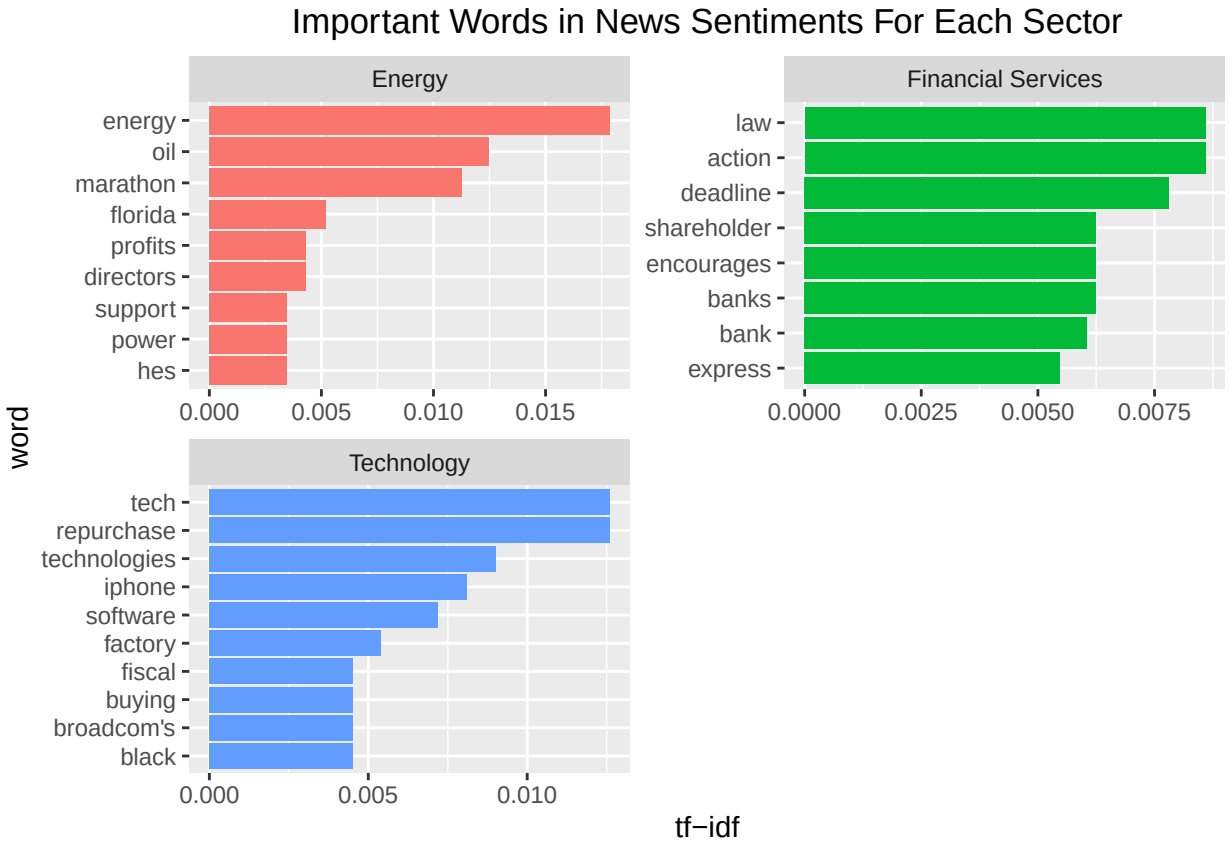

```
Pearson's product-moment correlation
```

data: proportion and Energy t = 10, df = 164, p-value <2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.5173 0.7063 sample estimates: cor 0.6207

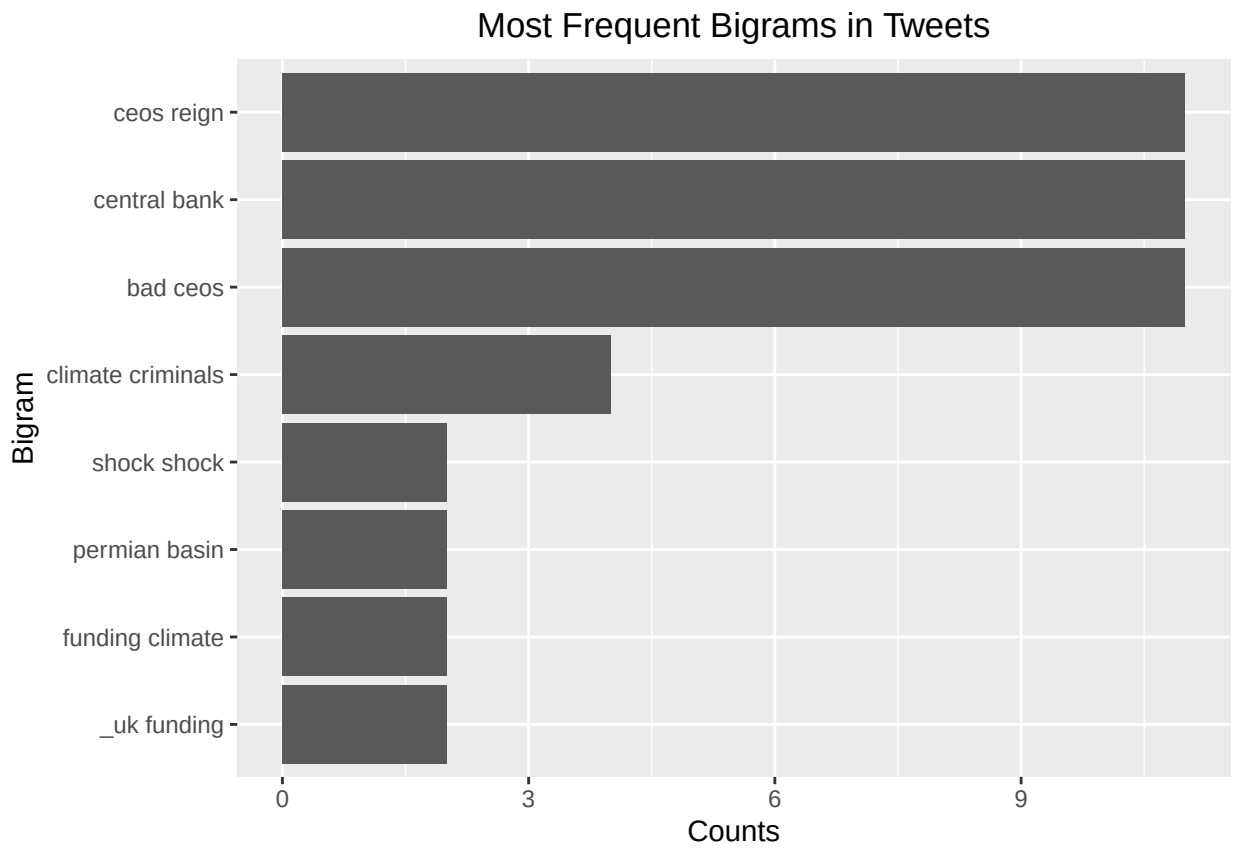
```
Pearson's product-moment correlation
```

data: proportion and Energy t = 20, df = 134, p-value <2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.8101 0.8991 sample estimates: cor 0.861

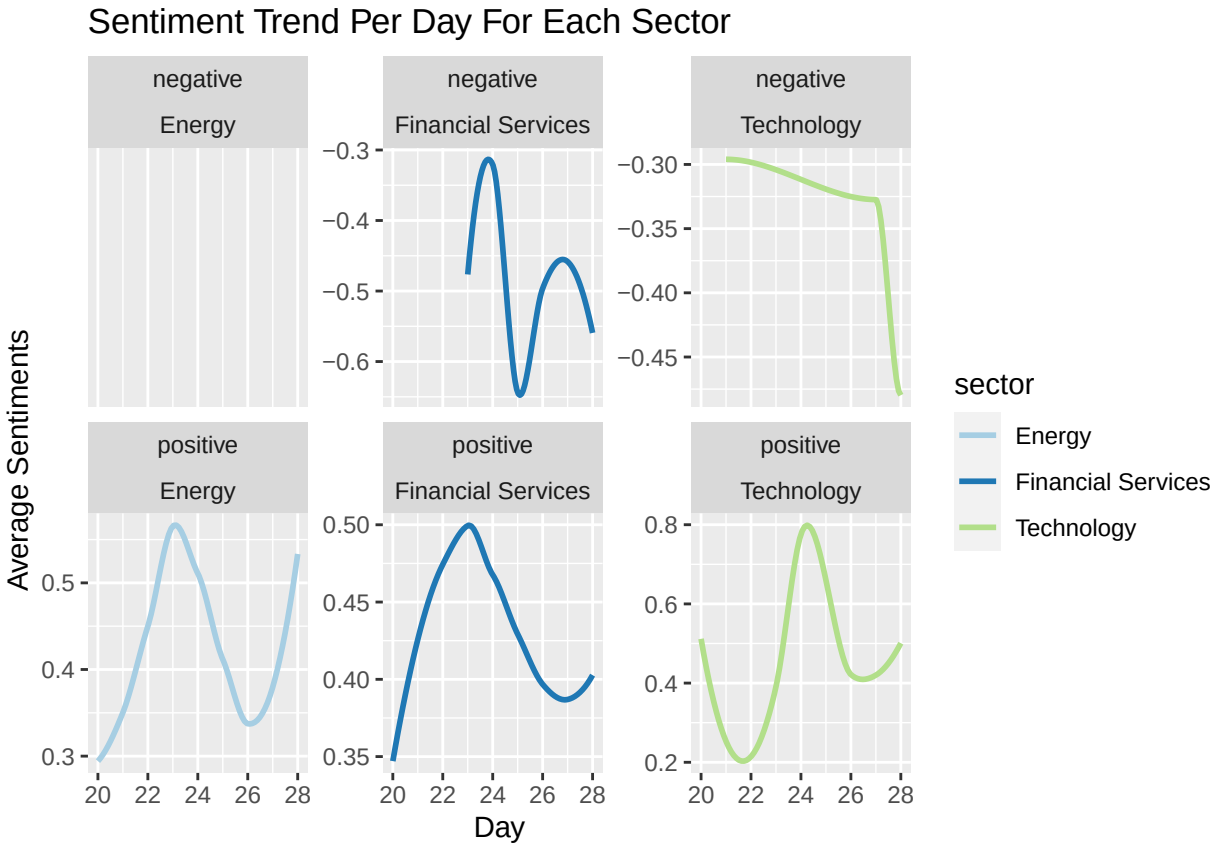
Since after comparing the three different sectors in terms of their News sentiments, it was found that the Energy sector had the most positive and least negative sentiments, we decided to compare the words used other sectors news to the words used in the Energy sector news. The words close to the line in the figure above similar frequencies in the news of both sectors. As it can be seen, words like market, investment, dividend, and accelerates had high frequencies in news for both Energy and Financial Services sectors, while words like accelerates, investors, betting, ahead, stock, among many others had high frequencies in news for both Energy and Technology sectors. Words like oil are strictly for Energy sector as expected, and words like bank are mostly used in the Financial Services line. One other interesting thing that can be seen from the chart above is that the news from the Energy sector had more common words in use when compared to the news from the Technology sector than when compared to the news from the Financial Services sector. The Pearson's product moment correlation values comfirms this; that computed for Energy aand Technology was 0.861, while the one computed for Energy and Financial Services was 0.6207.



From the chart above, it can be seen that the Technology Sector news consist mostly of words like tech, technologies, iphone, software which are words that are closely associated with the Technogy sector. And for Energy Sector, words like energy, oil, power appeared more often, which are also expected. The news from the Financial Services were centred around words used in that domain too, words like shareholder, bank, express, among many others.



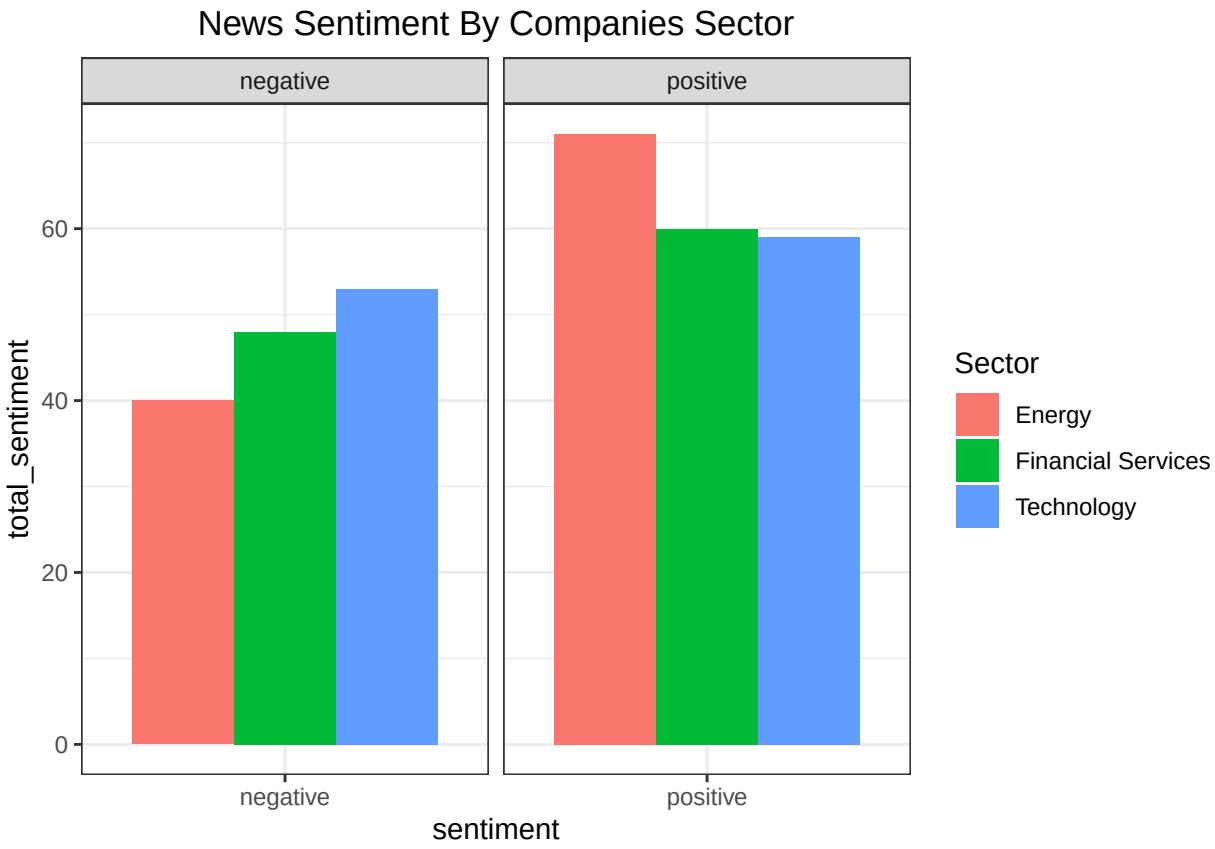
The chart above shows the most frequent bigrams used in tweets with negative sentiments from all the considered companies in this study. It looks like the actions of the CEOs of companies are the most responsible for making these companies have bad sentiments on social media. Also, it seems there are some criminal acts regarding fundings on climate in the UK that are causing negative sentiments on the news about these companies.



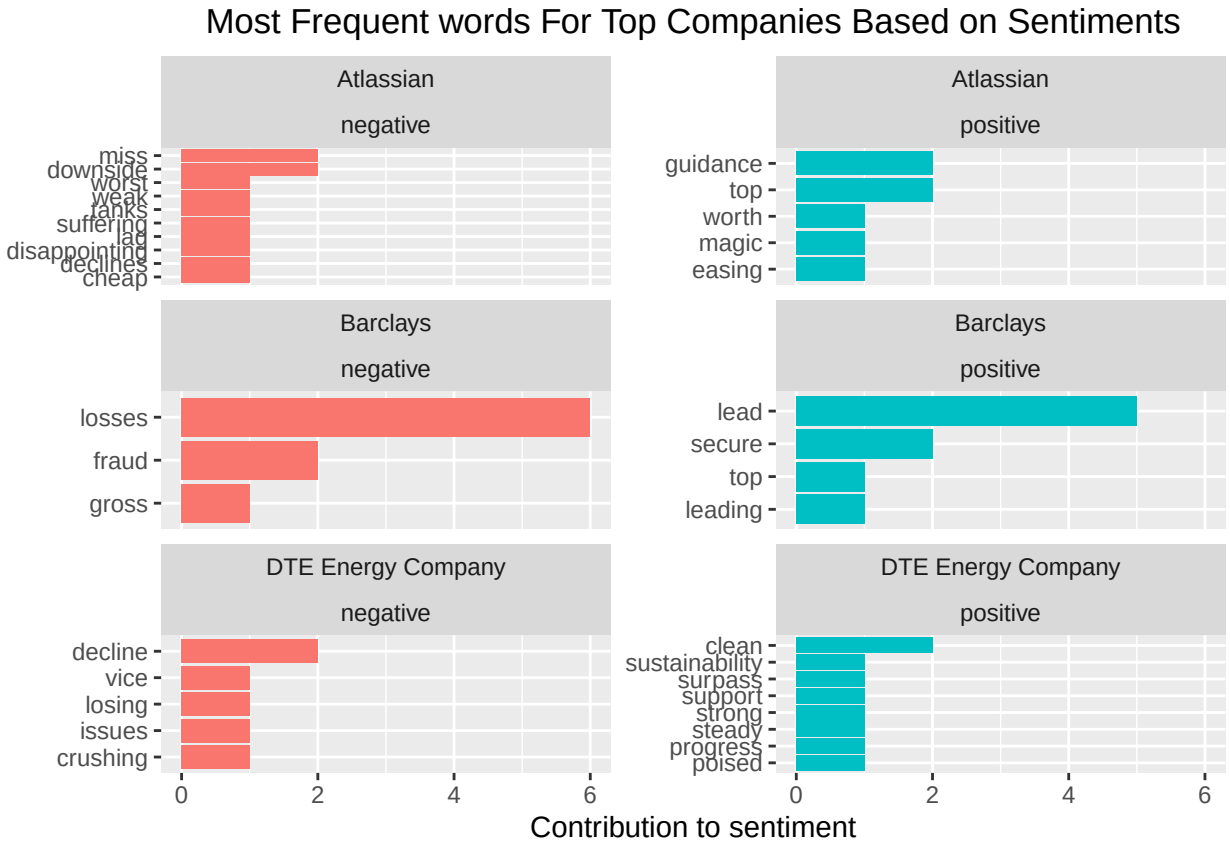
List of 1 $ plot.title:List of 11 ..$ family : NULL ..$ face : NULL ..$ colour : NULL ..$ size : NULL ..$ hjust : num 0.5 ..$ vjust : NULL ..$ angle : NULL ..$ lineheight : NULL ..$ margin : NULL ..$ debug : NULL ..$ inherit.blank: logi FALSE ..- attr(*, "class")= chr [1:2] "element_text" "element" - attr(*, "class")= chr [1:2] "theme" "gg" - attr(*, "complete")= logi FALSE - attr(*, "validate")= logi TRUE

It can be seen from the chart above that the Technology Sector looks like a no-go area for potential investors, the twitter sentiments around companies in this sector has been drastically increasing towards the negative end without any increase. Also, the Energy sector looks good, with maybe only 1 day or no day of negative sentiment, and constantly fluctuating amount of positive sentiments.

The chart above shows all of the relationships among words used in the tweets simultaneously, we particularly focused on tweets with negative sentiments. Firstly, looking into the network shaped as "3" in the chart, we could see that the chart linked up all words used in the tweets that were not English Language. Also, there were things around fire, live, army, and indian, which are associated with the ongoing war in Ukraine, these are mostly affecting the opinions about these companies as expected.



The chart above shows that the Energy sector had the least negative sentiment and the most positive sentiment, so the investor might want to consider investing in a company in these sector, but they should be wary about the recency of these tweets as well.



The chart above shows the most used words (for both positive and negative sentiments) in the tweets about the company with the most negative sentiment in each sector. It can be seen that Barclays is experiencing so much fraud activities. People are talking about how the Energy sector can be more sustainable and clean in their actvities, they are most likely making so much progress in these areas since it falls under the positive opinion of people.

Most Frequent Word in the Energy Sector by Sentiment

From the chart above, the most used negative words in this sector are crude, volatile, defensive, and falling, while the most used positive words are strong, support, and safe.

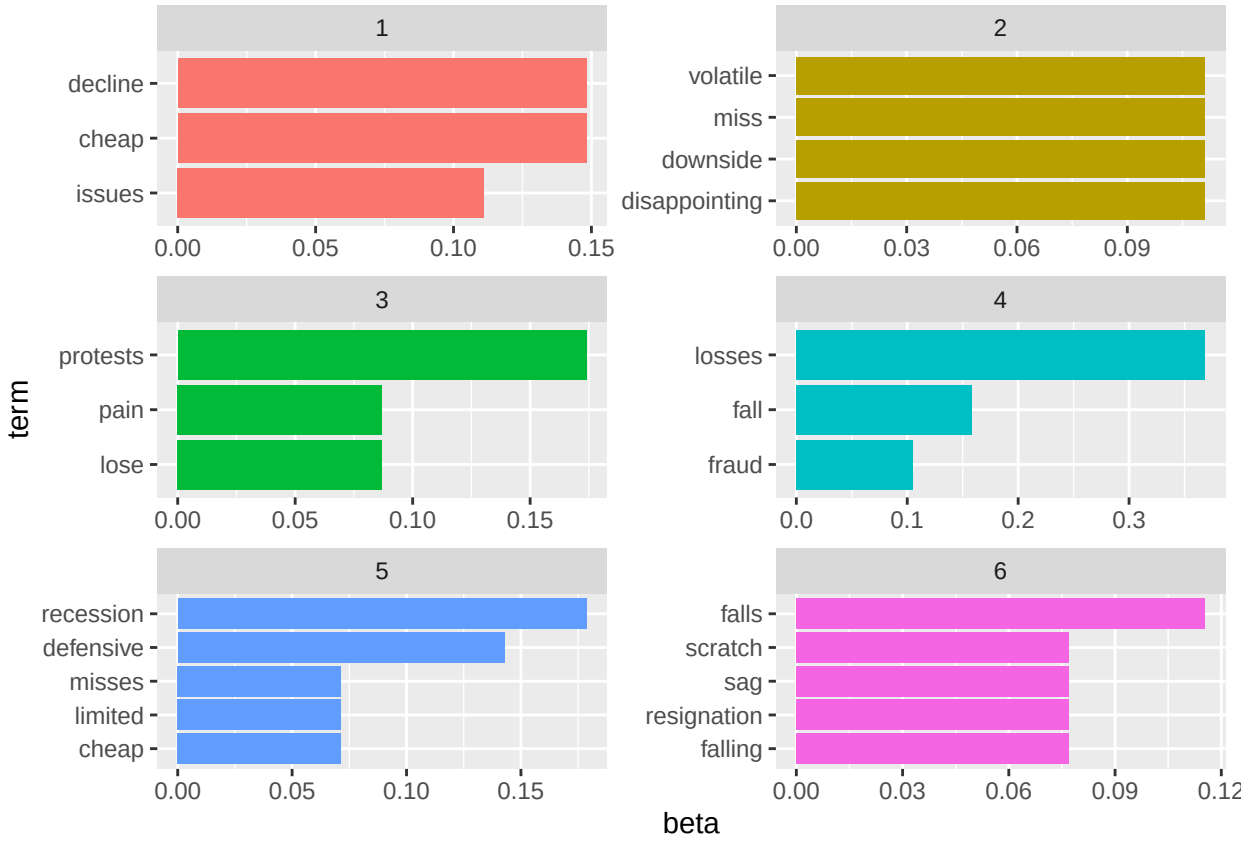
Most Frequent Word in the Financial Services Sector by Sentiment

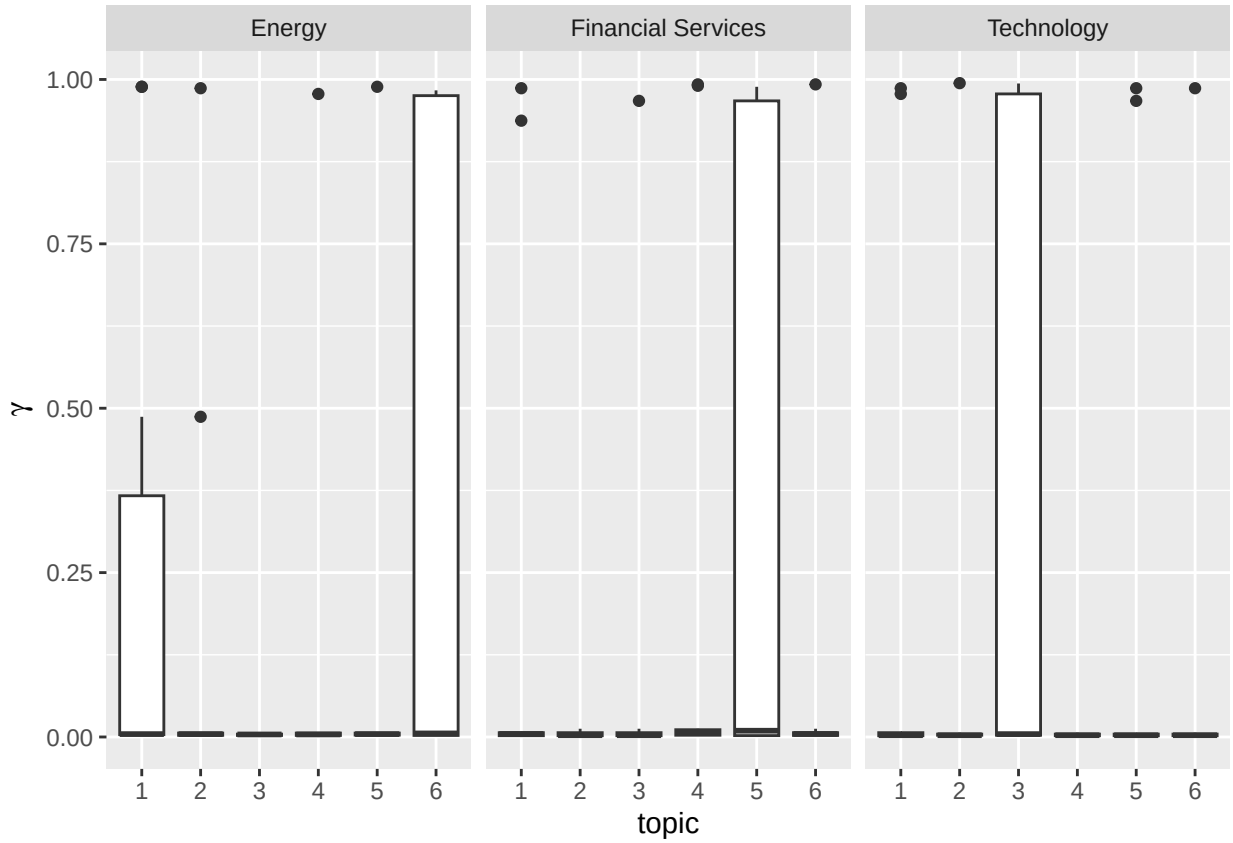
From the chart above, the most used negative words in this sector are losses, decline, and fall, while the most used positive words are top and lead.

Most Frequent Word in the Technology Sector by Sentiment

From the chart above, the most used negative words in this sector are protests, miss, issues, and downturn, while the most used positive words are top, ready, and guidance.

Topics of News with Negative Sentiments


1
decline
cheap
issues
0.00
0.05
0.10
0.15
2
volatile
miss
downside
disappointing
0.00
0.03
0.06
0.09
3
protests
pain
lose
0.00
0.05
0.10
0.15
4
losses
fall
fraud
0.0
0.1
0.2
0.3
5
recession
defensive
misses
limited
cheap
0.00
0.05
0.10
0.15
6
falls
scratch
sag
resignation
falling
0.00
0.03
0.06
0.09
0.12
term
beta



The focus here was on news that had negative sentiments, to see the issues that the companies are having. From the two charts above, it can be seen that the Financial Services Sector is solely about topic 5, which emphasizes recession. Also, the Technology Sector is solely about topic 3, which emphasizes protests and loses. Finallly, the Energy Sector is a mix of two topics, fairly topic 1 and mainly topic 6; talking about resignations, declines, issues etc.

Limitations/Next Steps

Limitations/Challenges Faced:

Bot Access/Term of Use: Some target websites for scraping data had very strict restrictions to scrape or allow any automation access. These websites were more rich in data and easier to scrape based on the HTML structure, but the studies had to be done with website that was friendly to web scraping.

Complicated Web Page Structure: The Web/html structure for the data source used for were complicated with no standard way of accessing some html tag content. Some of the easy ways of getting the xpath or css selectors (for example - SelectorGadget) didn't work out for such kinds of complicated web pages, so a thorough understanding of the pattern of the html structure had to be understood before manually coming up with the individual xpath for each data points.

Multiple Data Structure: All the data points needed were not on a single webpage which required creating multiple scrappers for each url and understanding the uniquely different structure for each url.

Quality of Data: The lack of quality data structure in the data coming from Stock Analysis and Twitter resulted in a lot of data manipulation and cleaning, they were highly unstructured in their natural form. For

the tweets, the usernames had to be removed, and even usernames with RT had to be removed, hashtags, and urls had to be removed from to get the clean texts that was used for analysis.

Ads Challenge: There were a lot of random Ad contents popping up when accessing the Stock Analysis website.

Limitations/Next Steps

More Sectors and Industries: For this study only 3 different sectors were considered, also only 10 organizations were selected in each of these sectors. This is limiting because there are many other sectors of companies that might be doing well and worth the investors interest, and even within each of the sectors there are tones of companies than the 10 just considered. So the plan is to include more sectors and industries within each sector, and maybe consider all the fortune 500 companies for a start and then later even more companies that didn't make fortune list but are on New York Stock Exchange (NYSE).

More Social Media Platforms: This study only included Twitter as the social mdeia source of data, there are some other popular and well used social media platforms that people often air their opinions about these companies that should be considered,

Tweets in other Languages: In this study, there was nothing done to cater for tweets about these companies written in languages different from English, and that was why the network graph group them together, so we couldn't get anything out of them. Moving forward, we would love to convert those language tweets into english first, before doing the required analysis on them, and these also has to be done accurately.

Conclusion

This Research shows that stock market price movement correlate with the public sentiments regarding the companies. And we could clearly see that the sentiment about the company in the media, industry reports, social media reviews or investors' opinions provided great insights into how the prices of stocks change, thus the texts and sentiments we gathered from people's opinion on social media, and important news on websites will enable investors learn more about the stock market and give them valuable insights that can be used to make investment decisions.

Our target audience for this project are strictly investors who are looking to make investments on stock or equity, they will be able to make use of our one-stop shop to get valuable information that can enhance their trading decision.

References

- https://www.tidytextmining.com/index.html
- https://youtu.be/MadMEVGMTUE
- https://github.com/aabeveridge/janitor_wrangle
- https://datascienceplus.com/parsing-text-for-emotion-terms-analysis-visualization-using-r-updated-analysis/
- https://towardsdatascience.com/social-media-analysis-802d7de085a3
- https://towardsdatascience.com/how-to-access-data-from-the-twitter-api-using-tweepy-python-e2d9e4d54978