Mining for Public Perception: Social Media and the News

Introduction

For the final project of Applied Data Mining, we were tasked with a very open-ended assignment: scour the data that the online world has to offer, and using the models that we have studied and used in class so far this semester, generate some non-obvious insights, findings, useful algorithms, and/or streamlined methods of handling and mining data to provide value for a certain audience. These valuable insights or tools that we come up with should offer increased saved time, decreased uncertainty, increased reward, or some other metric of value. Our audience should be somewhat targeted, and our insights/methods should be able to inspire further investigation and research by the target audience, or other data scientists seeking to develop/reproduce our methods and findings. This is equivalent to the "observation" step of the scientific method, where we systematically probe the vast data that the world has to offer to hopefully begin making valuable sense of all the noise and chaos.

Our initial thoughts naturally gravitated towards some of the more exciting and interesting datasets/sources we had been considering all semester. For example, Noah had seen a "Young People Survey Questionnaire" dataset on Kaggle which could have potentially been usable with other data to mine trends and craft methods for companies to market to young people, but upon further investigation, the data was not maintained, and the sourcing was a little unreliable. Coming to grips with the reality of data-reliability, our next project iteration saw us scouring the NYC OpenData platform as we did for Project 1, finding well-maintained data, but data that was often uninteresting for us, and unworthy of our final sustained focus of the semester. A NYC motor vehicle collisions dataset was interesting (especially when combined with a NOAA weather dataset) with plenty of features and details about car accidents, injuries, and deaths, but upon some exploratory data analysis (rough histograms of features, checking unique values of columns), interesting phenomena seemed few and far between, with aberrations to the norm happening extremely rarely, with these deviations poorly maintained in the dataset (extremely high NaN/None value counts, sketchy numbers for weather/crash statistics). We were facing an inspirational dead-end, and could not gain a sense of direction for a dataset we weren't very passionate about.

While peering over columns and columns of precipitation information in the weather dataset, we had an idea for something new. Inspired by a previous individual project that Noah had worked on over the summer and was in the process of building out, we began tossing around the idea of sentiment analysis, the use of natural language processing algorithms to analyze the meaning and intent behind text data. But what kind of sentiment analysis would offer useful insights? Well, the kind Noah had been getting familiar with the past summer on his own in Python: sentiment analysis gauging public and social opinion on companies, brands, and businesses. If we could gain some insights or develop

some methods pertaining to public sentiment polarity (how positive or negative the language is) and favorability of these institutions, the value would be clear with an obvious audience, and the models/methods in store for such an investigation would be abnormal (from the usual columnal datasets with tons of boring variables) enough to be engaging and dare I say it, fun. We got to work.

Brainstorming

Laying out an initial plan (which will be revised over time in this report, tracking the iterations of our ideas), we had the following ideas in place:

- Perform sentiment analysis on recent tweets containing reference to a specific keyword, in this case, a business name, to gain a sense of social media opinion
- Perform sentiment analysis on top news articles to gain a sense of public/institutional sentiment, a different type of polarity measure that could offer usefully-contrasting metrics and insights
- Develop a method for creating our own datasets regarding a specific business, one with tweets and their sentiment information, and one with news articles and their sentiment information
- Aggregate over these datasets to get summarizing sentiment information for a given business (polarity, subjectivity, keywords, etc), and engineer a feature pertaining to performance on social media vs. the news
- Compile an overarching dataset for each company to be joined on some larger dataset of top businesses, perhaps the fortune 500-1000 companies
- Analyze the resulting dataset, do keywords, industry, size, or other joinable factors (charitable donations, political involvement, corruption, etc) correlate or predict performance of our engineered sentiment metric?
- Validate our process with a potential deep dive into one stand-out industry, or even one company

After developing our methods and probing for insights, we would certainly uncover some interesting insights along the way, even if it would be difficult to distinguish sentiments in different sectors, or find important features for predicting public/social opinion. A lack of obvious findings would be a finding in itself- maybe corporate PR firms are misappropriating their efforts and resources? Either way, the process we envisioned would hit on all the points of a solid data-mining project and investigation, and it was time to get down to business.

The Data

The main data we are analyzing for this project is a bit unconventional in the sense that text from tweets and news articles do not carry the same conventional columnal data-table form, at least in the raw format that they “come” in. However, we will shape the data and our engineered features along the way into more familiar forms to work with.

Overall, the data sources are:

- n of the latest tweets containing a business name from Twitter's API cursor search
- No more than 25 of the top Google-News feed articles returned with a search query of the business name
- Fortune 2021's "Fortune 1000" dataset (downloaded from Kaggle)

The discussion below will outline (without going into code/package specifics, this is what the GitHub page will be for) how we accessed, prepared, and altered the data at each step, and in the rough order that we accomplished each step in order to maximize clarity of process, understandability, and reproducibility.

Twitter Data and News Media Data - Why?

Much of the motivation behind using these two different data sources is that, while sentiment regarding corporate dealings and entities within both sources are believed to influence markets, drive financial and economic trends, and perhaps even have a hand in dictating the success or failure of different ventures, they differ in some key insightful ways. While twitter may often offer an unfiltered view of non-human institutions and entities with a refreshingly informal honesty, reliability, and transparency direct from actual people, news articles may offer a complementary assessment of more institutional, academic, corporate, or political perspectives regarding some business or entity. Together, we feel that these two data sources complement each other, each filling in the shortfalls of the other to try to cover a very broad spectrum of sentiment that may be insightful when analyzed.

It is hard to imagine emerging, influential conversations that could drive any sort of trends that a business might care about not happening, at least in some part, on Twitter, or on articles rising to the top of Google News. A shortcoming that is destined to impact a company's success may begin festering online in tweets and articles long before manifesting in the marketplace, and likewise, the embers of success and profitability may be recognizable in the sea of text data being produced on social media and news media before any PR or marketing interns notice them.

Additionally, while news article data that we will look for (ie, the top articles returned by Google) are only "produced" by "nature" every so often (a handful per day), tweets provide a constant stream of information that is always up to data and ever-changing at every moment. In this sense, a combination of the two may be able to better capture both fleeting and lasting trends worthy of note.

Either of these datasets could be considered large or small depending on your perspective. Several tens of news articles may contain thousands of words, which really isn't all that much text data in terms of sentiment analysis and NLP tasks, but a practically unlimited number of search results related to a query in some way are available at the click of a button, or setting of a parameter. Likewise, with the tweet data, thousands of tweets may seem like a lot, but really might not

even extend back before the current day depending on the query. For both data sources, we will experiment with both "large" and "small" calls/queries to see if there is much of a difference. In the end, we believe that tending on the smaller side will benefit our devices and storage capabilities, and won't be too impactful for the meta-results we hope to find when extending the process over a long list of companies, each having hundreds of tweets and tens of full articles analyzed for them. Now, let's dive into the data and look around.

Twitter Data

Like we explored in class, once you make a connection to Twitter's API, you can practically query tweets programmatically with the same effectiveness as logging onto Twitter and using the "Explore" feature (our data source is Twitter in general, an entity that does maintain its data). This would be important for us, as in order to gather the data to be mined, we would need to search the Twitter timeline for recent tweets containing a given keyword, which is in this case, a company name, like "Microsoft". With some set parameter n-number of tweets, we will have a text dataset to begin performing sentiment analysis on.

Twitter Data - Packages (Python)

- **API access**
  - Tweepy
  - Requests
- **Data formatting and cleaning**
  - Pandas
  - Numpy
  - Re
  - String
- **Sentiment Analysis**
  - NLTK
  - TextBlob
  - Newspaper3k

Twitter Data - Loading and Cleaning

With our API credentials loaded, we query the n most recent tweets containing our keyword. The first thing we notice upon analyzing the output is that there are many duplicate tweets. We don't want duplicate tweets making it into the final dataset, as this could skew the results by artificially increasing the frequency/presence of a certain sentiment, so we drop duplicates from our returned set. It is important to note here that the eventual target of 250 tweets per company name stretches different amounts of time for different companies. 250 tweets for *Google* might stretch a few minutes, while 250 tweets for a company like *Plains GP Holdings* might stretch a few months. This will definitely affect how closely the tweets relate to the news articles in the future, or even how they relate to the Fortune 1000 data, sacrificing how up-to-date our information is.

The next thing we notice, like in our in-class exploration, is that the tweets themselves contain some meaningless content that would certainly not contribute to the sentiment of the tweet, like the "RT" for retweets, the "@username" at the beginning of replies and quotes, etc. We replace all of these unneeded string characters using regular expressions, and some normal parsing.

We make the "executive decision" to initially leave in punctuation and numbers, as the packages we will use to perform sentiment analysis have pre-built systems for formatting the input text it receives, and things like large or small numbers, or various punctuations like question marks and exclamation points may actually contribute to the polarity and subjectivity. Additionally, the methods we will use to compile keywords also strip and format appropriately.

In the same vein, tokenization and lemmatization is left to the pre-compiled sentiment analysis and keyword determination packages that we use, and we would like to be able to make more sense of the keywords upon them being spit out to us and being present in our dataframe.

Twitter Data - Sentiment Analysis and Keyword Generation

Once we have our text data cleaned and formatted (and put into vectors/series), we run sentiment analysis on each tweet. We compile the tweet's polarity (which we can use to assign a label "positive," "neutral," or "negative") and subjectivity. These metrics will come in handy when we go to engineer features later, and interpret the results of our mining.

We also do what we can (utilize NLP capabilities of our packages) to gather keywords from each tweet, and save those as well, as they will also come in handy when we attempt to interpret our results, and perhaps run alternative models in the future.

Twitter Data - Export

With all of these fields populated for each tweet, we construct a dataframe for the roughly n (some less due to duplicates removed, can be forced to be n optionally) tweets. This is the output of our twitter data processing function that uses parameters *word* and *n*, and returns this dataframe of *text, polarity, subjectivity,* and *keywords*. We will eventually aggregate over these dataframes we can produce for each company to fill a company dataset.

News Article Data

With a little bit of research, and some previous knowledge from Noah's past project, we found that streamlined methods exist to return news articles from the top pages of Google News (our data source) from a given search query, and then their content could be parsed. With a query like "Microsoft" (and language set to "en"), we can return all of the first page news results Google would return about Microsoft. This is important, as we can compile some number n of these top articles as the data to run sentiment analysis on, and gain a sense of the more public, institutional, and political sentiments surrounding a business.

News Article Data - Packages

- **Article Access**
  - PyGoogleNews
  - Newspaper3k
- **Data Formatting and Cleaning**
  - Pandas
- **Sentiment Analysis**
  - TextBlob
  - NLTK
  - Newspaper3k

News Article Data - Loading and Cleaning

With PyGoogleNews capable of querying links to Google News results articles, and Newspaper3k capable of taking these links and parsing the articles that they correspond to, the job of loading the text data for analysis was straightforward. Cleaning and formatting was also relatively short for the news article data. As the packages we use for sentiment analysis and keyword generation were designed to run on raw and scraped text sources from the internet (and other places), they contain their own "under the hood" preprocessing, tokenization, and lemmatization systems that we decided not to overwrite/reimplement. We did not find any issues with resulting analyses and keywords later, so we decided that keeping this part simple would be better. All we did was process the fields *title, text, keywords, polarity,* and *subjectivity* into a neat dataframe to be filled.

News Article Data - Sentiment Analysis and Keyword Generation

Once we have the text from each top article, we run sentiment analysis on the article and collect the polarities, subjectivities, and keywords, just like we did for the tweet data. Again, we will be using these metrics to engineer features in the future, draw conclusions/insights from our results, and maybe run even other models on these features in the future. Like in the tweet data, we also gather keywords for each article that can potentially provide insights and relate to collected sentiments. These things are all compiled into lists and series for export.

News Article Data - Export

With all of the fields populated for each news article, we construct a dataframe of the top ~20 articles and their relevant information compiled in the previous steps (note that this number could be higher with higher processing power, more storage resources, ie, a corporate/industrial setting), and now this data is ready to be saved as csv, or aggregated into a larger dataset, as we will see occur.

Forbes Fortune 1000 Data

Now that we have a process that can be run for a company to gather its relevant features that we discussed above from the tweet data and news article

data, all we do is iteratively perform these two steps on each business name present in the Forbes Fortune 1000 dataset. As we output dataframes of tweet analysis and news article analysis data for each company, we aggregate over the columns to compile the following information to be joined into the company dataset

- Average polarity from articles
- Min polarity, as well as min polarity article info (title, text, keywords, etc, can be scaled to have min k polarities)
- Max polarity, with info
- Average polarity for tweets
- The above max and min for tweets
- Average subjectivity for tweets, articles
- Most frequent/significant keywords from tweets, articles

The dataset includes fields that we will carry along to be potential components in engineer features, or raw features themselves: *rank, sector, industry, city, state, num_employees, ceo_woman, profitable,* etc. This data is available at the link: https://www.kaggle.com/winston56/fortune-500-data-2021?select=Fortune_1000.csv . Overall the dataset contains very few empty values, which we will fill with numerical means where appropriate, and merely ignore when they are categorical, and we don't want the nulls/nans to create errors in our eventual algorithms. It is a Kaggle dataset, but since it is a third source of data really only necessary for the company names and some basic information, we felt that it was acceptable to use. The dataset has updated Fortune 1000 information through the end of February 2021, so we will consider it up-to-date.

Overall Company Dataset - Joins

Once we have aggregated the tweet and article data as mentioned before, and we have our Forbes Fortune 1000 dataset to work, we simply perform a join on the company name/search query, filling an overall company dataset with all of our fields and variables we have discussed along the way. This will be the main dataset we use going forward. Note that we only use the top 100 companies due to processing and storage constraints, but in an industry setting we would have the resources to build out our method and use a larger dataset, if we aren't already isolating it to a single company for specific research purposes.

Overall Company Dataset - Feature Engineering

Aside from some of the obvious features that we have engineered and merged into our dataset like sentiment analysis information, keywords, etc, we must return to one of the central motivating factors behind this project, which is the quantitative determination of a company's recent "performance" in the social media and news media sectors, in terms of public sentiment. To explore this fully, we see an excellent opportunity to engineer some features that may provide

useful insights in the end. Here is a list of the engineered features, some trivially simple, others less so:

- Average polarity (avg polarity score = (twitter polarity + news polarity)/2)
    - Will serve as a single number indicator of overall polarity for a company, ie, how do people and institutions feel in general?
- Average subjectivity (calculated analogously as above)
    - Will serve as a single number indicator of overall subjectivity for text discussing the company, how opinionated versus objective are people when they discuss the company?
- Media Sentiment Difference (Twitter average polarity - Article average polarity)
    - Will serve as a score that measures performance on social media versus news articles, a negative score is better performance in the news, positive is better performance on social media
- Media Sentiment Intensity (|Twitter average polarity| + |Article average polarity|)
    - A single number indicator of how intense the sentiments are for a company overall. A higher number means greater intensity (for both positive and negative), lower means more neutrality.
- CEO Gender
    - Using NLP to predict on CEO gender from CEO name, similarly to how this was done in lecture.

Here is what (a part of) the final table looks like:

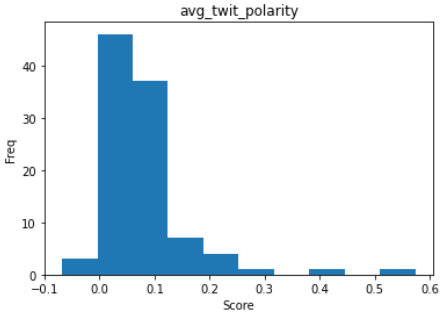
	company	rank	rank_change	revenue	profit	num. of employees	sector	city	state	newcomer	...	max_twit_polarities	max_twit_text	max_
0	Walmart	1	0.0	523964.0	14881.0	2200000	Retailing	Bentonville	AR	no	...	0.5 0.8 0.5	walgreens: bellevue, blair, columbus, fremont,...	0.209 0.25
1	Amazon	2	3.0	280522.0	11588.0	798000	Retailing	Seattle	WA	no	...	0.65 0.7 1.0	rt @rishaaaaaaa_: streaming on spotify and ama...	0.330 0.413
2	Exxon Mobil	3	-1.0	264938.0	14340.0	74900	Energy	Irving	TX	no	...	0.6 0.75 0.8	rt @drafzalniaz: the four horsemen of banking ...	0.15 0.167
3	Apple	4	-1.0	260174.0	55256.0	137000	Technology	Cupertino	CA	no	...	1.0 1.0 1.0	i'm gifting you doosra card for paytm cricket ...	0.26 0.354
4	CVS Health	5	3.0	256776.0	6634.0	290000	Health Care	Woonsocket	RI	no	...	0.6000000000000001 0.625 0.875	this job might be a great fit for you: pharmac...	0.218 0.230
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
95	Northrop Grumman	96	12.0	33841.0	2248.0	90000	Aerospace & Defense	Falls Church	VA	no	...	0.5 0.5 0.75	an overwhelming majority of americans favor en...	
96	Capital One Financial	97	1.0	33766.0	5546.0	51900	Financials	McLean	VA	no	...	0.9 0.9 1.0	proud to announce that capital one was named a...	0.21 0.223
97	Plains GP Holdings	98	-4.0	33669.0	331.0	5000	Energy	Houston	TX	no	...	0.0 0.0 0.0	$pagp / plains gp holdings files form def 14a ...	0.129 0.167

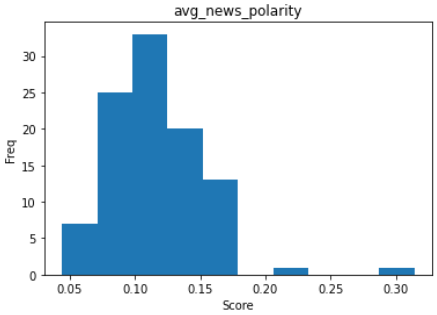
Data Exploration and Mining

Once our company dataset was put together, the first thing we wanted to do was some very basic exploratory data analysis. What is the range of values that our sentiment analysis variables and engineered features take on? Can we unearth anything initially through more naive ways of analyzing data? Let's find out. Throughout this section, bolded sentences reflect the seeds of potential insights and value which will be discussed at the end of the report.

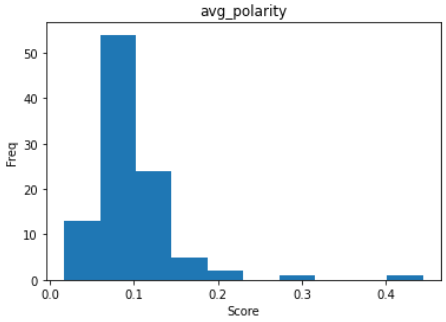
Variable Distributions

- Polarity

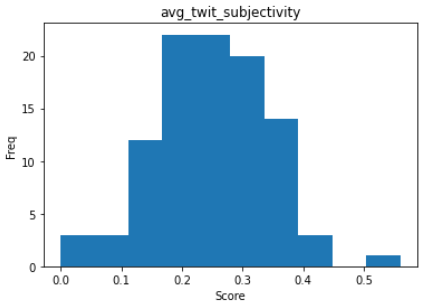


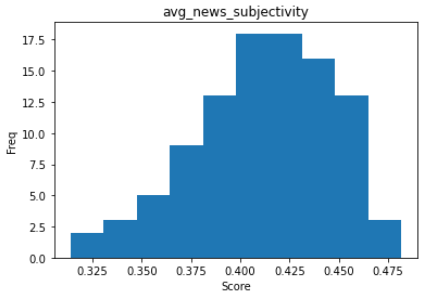


- Average Overall Polarity - Polarity Average from tweets and articles (weighted equally)

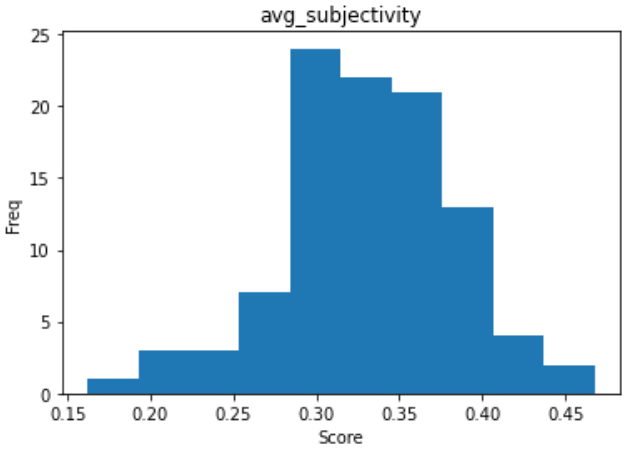


- Subjectivity

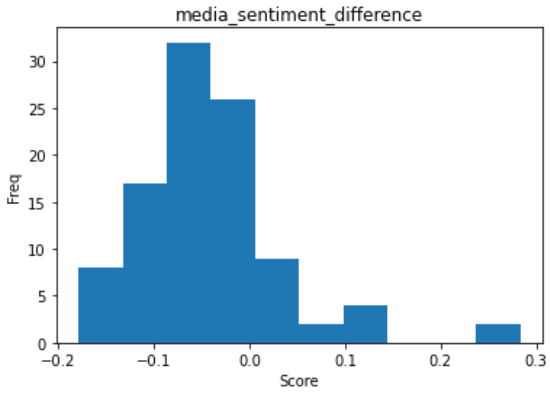


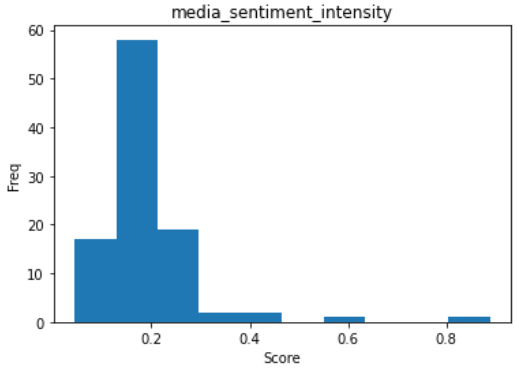


- Average Overall Subjectivity - Average from tweets and articles (weighted equally)



- Media Sentiment Difference & Media Sentiment Intensity





Before discussing the relative spreads, we will briefly comment on the data condition. There were only a couple (< 3) companies that we could not return enough tweets or articles to create polarity/subjectivity scores for, and in these cases, we merely returned None values that can be explicitly ignored or filled in later. This is a benefit of using the Fortune 1000 companies - there is a lot of online information about them, so there are limited empty fields that we see while developing these datasets.

From this initial data exploration, we see that the polarity scores (positive vs negative sentiment) for tweets and for news articles both seem to be centered close to .1, and slightly skewed right. However, the **news polarities are noticeably a bit more spread out from .05 to .2, while the tweet polarities are more compact around .1**. Perhaps this suggests that **the way that social media users consume and write media is the same for large institutions in general**, while news outlets'

intentional hyper-focus on the specifics of a certain company **may be reflected in more spread-out distribution of news polarities** for different companies. We find that truly negative average polarities seem to be extremely rare overall, so that is something to consider going forward. Overall, the combined polarity average does not tell us much that the separate averages do not already.

Subjectivity for both tweets and articles both seem to be close to normally distributed, with news subjectivities being a bit skewed left. Their average looks slightly skewed left as well. Something that is unexpected however, is that **the center of the tweet subjectivity distribution is actually a noticeable bit lower than the news subjectivity center**. This means that, for this sample of data collected, the **news articles used actually more subjective language than the tweets did**. This is an interesting finding, and while we will discuss it more in depth later, it is worth noting that this may be a truly useful insight, reflecting the potential truth that **social media users speak very objectively when they send their subjective opinions out into the void, while article writers understand their target audiences, and their impact quite well, so their language is a bit more targeted and persuasive**, rather than purely objective, as one would expect of good journalism. However, this also may be a product of a wider range of vocabulary from the articles making into the sentiment analysis of the text for each article compared to each tweet, but this possibility is discredited by the closeness of the polarity distributions, with only a slightly larger spread for news articles compared to the relatively large gap in centers for the subjectivity distributions.

Next, what we believe to be the most important feature, media sentiment difference- a score capable of assessing a company's relative sentiment on social media versus the news. While we will break down this metric in more ways, for now, we see that it is highly centered around -.1 to 0, suggesting **that most of these companies are doing slightly better in the news than in tweets**. Media sentiment intensity doesn't tell us much yet, and merely reflects the sum of the polarity distributions.

Let's begin applying some models and algorithms to further investigate the data.

Correlation

One of the most useful algorithms that we feel can unearth interesting insights for our data happens to be one of the simplest and most widely used: correlation. If we can find even slightly non-trivial correlations in our data, it may either confirm certain speculations about the media sphere, or lead us down some paths for more investigation, or inspire future research. This is the worst case- but in the best case, perhaps a stronger correlation than expected can be found and provide value in itself.

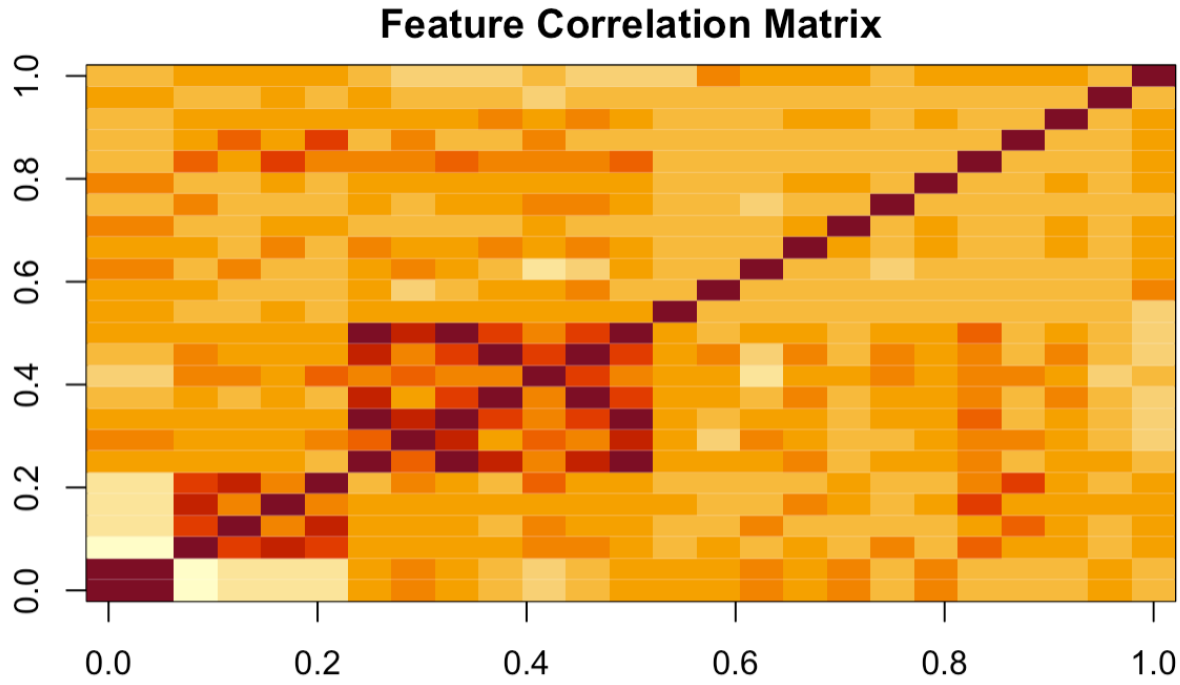
First, we need to subset out the columns in the dataset that we feel are relevant to use in a correlation matrix. We create a dataframe of all of the numeric features, and insert dummy variables of 0s and 1s to fill in for different industries, as well as male and female CEOs. Additionally to preprocess the data, we filter out industries that have three or less companies of representation in this top 100. Also, in terms of data quality, some of the success indicators like profit, market

cap, etc, had some (3-4 total) missing values that we filled in with means to smooth over the correlation process. The final numeric data frame looks like this (the dummy variables are not visible, they're at the far right of the table):

	rank	revenue	profit	num_employees	mkt_cap	avg_twit_polarity	avg_news_polarity	avg_polarity	avg_twit_subjectivity	avg_news_subjectivity	m
0	1	523964.0	14881.0	2200000	411690.0	0.041710	0.109596	0.075653	0.217693	0.354070	
1	2	280522.0	11588.0	798000	1637405.0	0.031813	0.169450	0.100632	0.128812	0.454728	
2	3	264938.0	14340.0	74900	177923.0	0.079331	0.072307	0.075819	0.338149	0.444209	
3	4	260174.0	55256.0	137000	2221176.0	0.032285	0.159517	0.095901	0.145171	0.447284	
4	5	256776.0	6634.0	290000	98496.0	0.261230	0.122886	0.192058	0.446915	0.469504	
...	...	...	...	...	...	...	...	...	...	...	
95	96	33841.0	2248.0	90000	49812.0	0.044201	0.162379	0.103290	0.141154	0.449329	
96	97	33766.0	5546.0	51900	50946.0	0.212082	0.100178	0.156130	0.310661	0.350793	
97	98	33669.0	331.0	5000	1463.0	0.000000	0.044031	0.022015	0.231250	0.417530	
98	99	33266.0	7882.0	30000	198828.0	0.045278	0.088460	0.066869	0.192039	0.435170	
99	100	32897.0	85.1	2012	988.0	0.048810	0.140223	0.094516	0.344048	0.386466	

100 rows × 14 columns

We create a correlation matrix for every numeric feature and these dummy variables. Here is the corrplot for it:



We will highlight some of the correlations that we believe are interesting, and worth further investigation. In the bullets below, note that most of the clearly uncorrelated columns had correlation coefficients between -.1 and .1, if not even less strong.

Sentiment and Sector

- Media Sentiment Difference with:
  - Retail Sector: .21
  - Technology: -.22
- Average Polarity with:
  - Retail Sector: .29
- Twitter Polarity with:
  - Retail Sector: .25
- News Polarity with:
  - Energy Sector: -.29
  - Retail Sector: .24
- News Subjectivity with:
  - Technology Sector: .21
  - Financial Sector: -.44
  - Transportation Sector: -.31

Sentiment and Business Variables

- News Subjectivity with:
  - Revenue: .18
  - Market Cap: .27
- News Polarity with:
  - CEO Gender Male: -.3
- Twitter Polarity with:
  - CEO Gender Male: -.2

Sentiment and Sentiment

- Twitter Polarity with News Polarity: .4
- Twitter Subjectivity with News Subjectivity: .2

Interpretation

These above correlations were all of the noteable correlations that came out of the correlation matrix. While the numbers don't seem to be that high, it is worth remembering that all of these variables are "drawn" from completely different sources, and without performing any inferential statistics, we believe that these numbers, especially those >.25, represent some significant linear relationships, and is indicative that the correlation algorithm is fitting the data somewhat meaningfully.

Firstly, our favorite feature, media sentiment difference, seemed to be mildly correlated with the traits of being in the retail sector and the technology sector. We see that with the presence of a **retail company, media sentiment difference trends upward, indicating retail companies do slightly better in the news than on twitter, while the difference trends downward with the presence of technology companies, indicating better sentiments on Twitter**. We can actually note that the retail sector is weakly correlated with polarity score in both twitter and the news.

We see that polarity score and the energy sector are somewhat correlated negatively, indicating **"negative" sentiments in the news for the energy sector**.

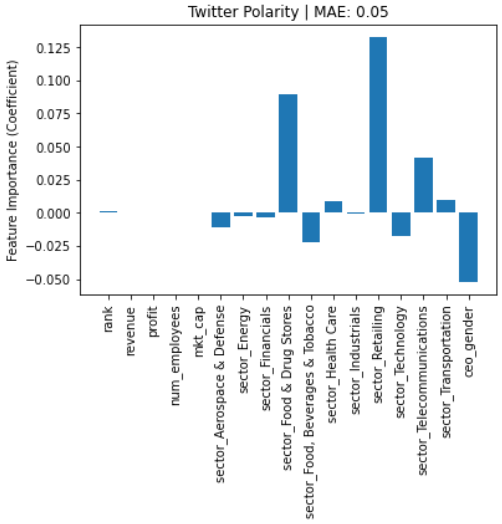
However, moving to news subjectivity, we find perhaps the strongest correlation: the financial sector with the news subjectivity score (-.44). **The presence of a company in the financial sector actually appears to be inversely correlated with news subjectivity, indicating that more objective news reporting is done for the financial sector.** Comparing this finding with the correlation between twitter subjectivity and financial sector, which wasn't listed as significant (-.1), it is interesting to note that **in news articles, there appears to be somewhat of a structured relationship between financial sector companies and subjectivity, but this structured dynamic does not appear to exist on Twitter.** We also see a comparable **negative correlation for the transportation sector, and a weaker positive correlation for the Technology sector.**

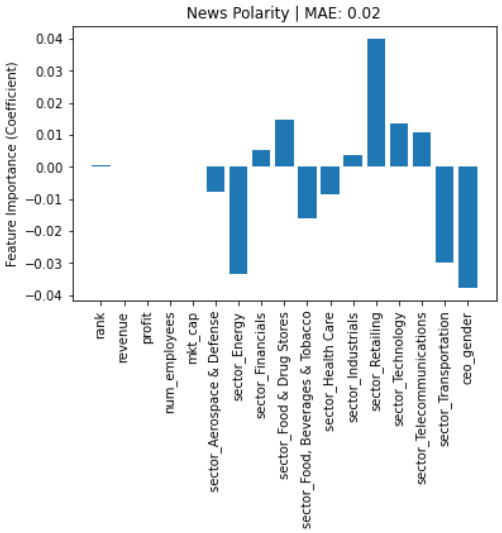
Moving to correlations between sentiment variables and business variables, we note a weak correlation between market cap and news subjectivity, indicating **more subjective language is present in news articles for companies with larger market caps.** Even more interesting is the negative correlation between a company's CEO being male, and news subjectivity. This means that **news article language tends to be more subjective for female CEOs, and more objective for male CEOs**.

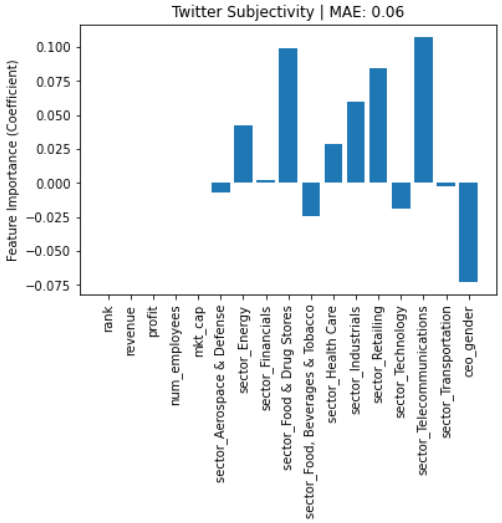
Finally, an interesting correlation between two of the sentiment analysis variables- we see that **Twitter polarity score and news polarity score are in fact mildly positively correlated**. We see that the two scores do not trend opposite from each other, and can even be said to somewhat increase and decrease together, but perhaps not as much as expected, which we will discuss again later in a more formal analysis section.

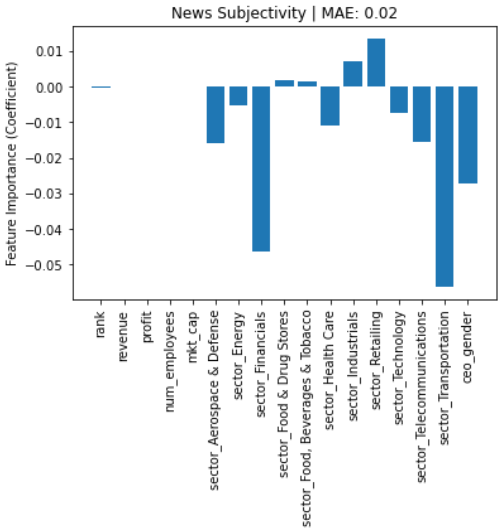
Regression Models for Polarity and Subjectivity

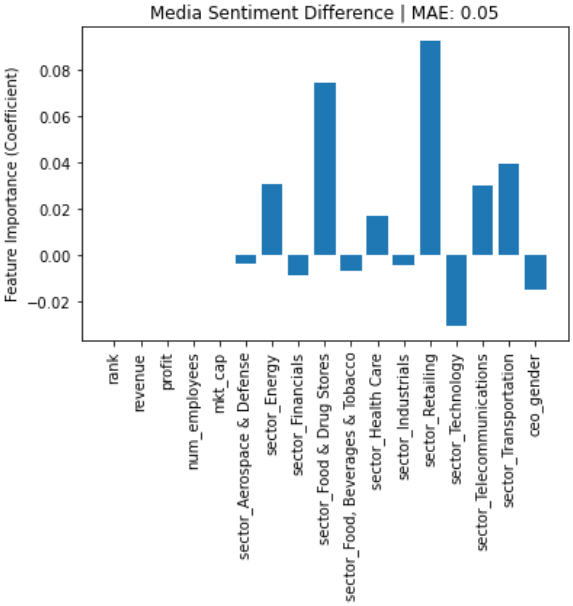
With the results from the correlation investigation in mind, we want to see if linear models for polarities and subjectivities can be constructed from the other features in our dataset. This was the next logical step that we had in mind if the correlations returned a few non-trivial results, which we believe it did. We will use ordinary least squares (OLS) to attempt to fit a regressor and analyze the coefficients/important features. We will also calculate the MAE, or mean average error, that each OLS model has in fitting the target variable. We will fit this on the entire data as the "training" set, as we are not trying to develop a predictive tool, and merely want to see how well an OLS model fits all of the data we have. Perhaps these processes will either verify some of our findings from the correlations, or maybe steer us in a different direction.


Twitter Polarity | MAE: 0.05
Feature Importance (Coefficient)
rank
revenue
profit
num_employees
mkt_cap
sector_Aerospace & Defense
sector_Energy
sector_Financials
sector_Food & Drug Stores
sector_Food, Beverages & Tobacco
sector_Health Care
sector_Industrials
sector_Retailing
sector_Technology
sector_Telecommunications
sector_Transportation
ceo_gender


News Polarity | MAE: 0.02
Feature Importance (Coefficient)
rank
revenue
profit
num_employees
mkt_cap
sector_Aerospace & Defense
sector_Energy
sector_Financials
sector_Food & Drug Stores
sector_Food, Beverages & Tobacco
sector_Health Care
sector_Industrials
sector_Retailing
sector_Technology
sector_Telecommunications
sector_Transportation
ceo_gender


Twitter Subjectivity | MAE: 0.06
Feature Importance (Coefficient)
rank
revenue
profit
num_employees
mkt_cap
sector_Aerospace & Defense
sector_Energy
sector_Financials
sector_Food & Drug Stores
sector_Food, Beverages & Tobacco
sector_Health Care
sector_Industrials
sector_Retailing
sector_Technology
sector_Telecommunications
sector_Transportation
ceo_gender


News Subjectivity | MAE: 0.02
Feature Importance (Coefficient)
rank
revenue
profit
num_employees
mkt_cap
sector_Aerospace & Defense
sector_Energy
sector_Financials
sector_Food & Drug Stores
sector_Food, Beverages & Tobacco
sector_Health Care
sector_Industrials
sector_Retailing
sector_Technology
sector_Telecommunications
sector_Transportation
ceo_gender

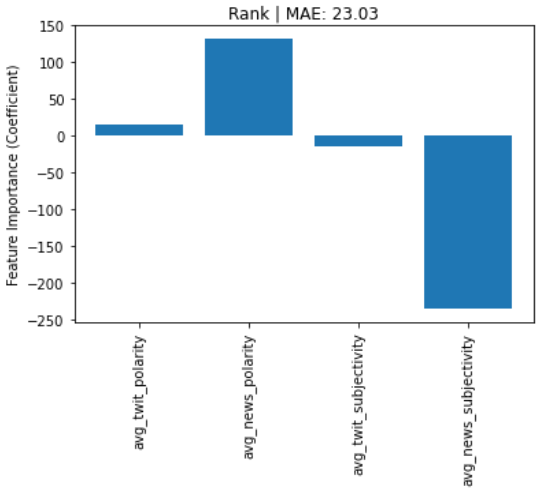

Media Sentiment Difference | MAE: 0.05
Feature Importance (Coefficient)
rank
revenue
profit
num_employees
mkt_cap
sector_Aerospace & Defense
sector_Energy
sector_Financials
sector_Food & Drug Stores
sector_Food, Beverages & Tobacco
sector_Health Care
sector_Industrials
sector_Retailing
sector_Technology
sector_Telecommunications
sector_Transportation
ceo_gender

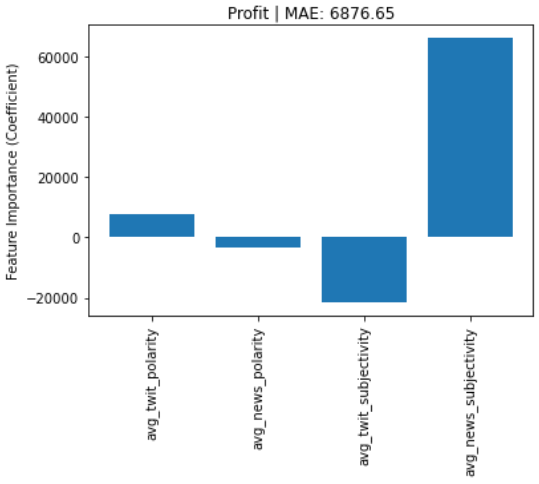
First looking at Twitter Polarity, we see that **the Retail variable has the largest overall coefficient, and is very positive, with Food and Drug close behind**. Telecommunications is small and positive, and ceo-gender-male is small and negative. This further indicates that **there exists a direct relationship between retail companies and polarity**, in that a company having the trait of being in the retail sector also trends more positively in tweet polarity, and vice versa. **The same for the retail sector can be said for news polarity more so than any other sector**, with transportation, energy, as well as ceo-gender-male having noticeable negative coefficients. **The energy sector and ceo-male negative relationships were also spotted in the correlations, lending further credence to these relationships**.

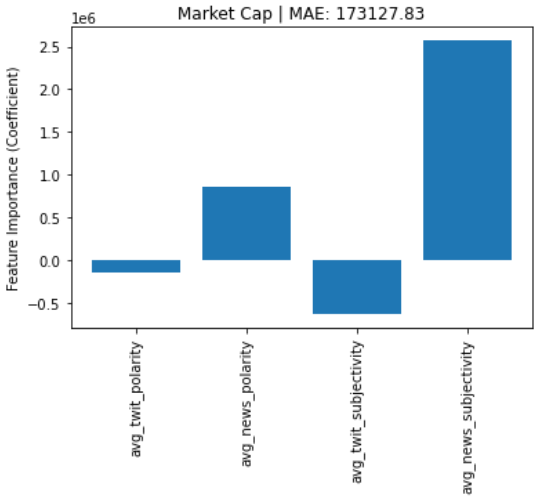
For tweet subjectivity, we see that there are a few potentially significant positive variables, like food and drug, retail, and telecommunications, with ceo-gender-male having a negative coefficient. Twitter subjectivity relationships weren't really spotted in the correlation analysis, but moving to news subjectivity, we see that **the financial sector and transportation sector have sizeable negative coefficients, indicating news subjectivity's negative relationship with these variables.** These negative relationships were also picked up in the correlation analysis, further justifying them.

Finally, taking a look at media sentiment difference, **the food and drug variable, and the retail variable appear to have noticeable positive coefficients**. The correlation analysis picked up the retail relationship, but also suggested a negative correlation with **the technology sector, which does have a somewhat negative coefficient there as well.**

These regression results made us want to look at some of the business metrics, and see if OLS could identify some of the key variables for things like rank, profit, and market cap.







The most significant variables in the rank regression were both news polarity and news subjectivity, with polarity having a positive coefficient, and subjectivity having a negative coefficient. This would suggest that rank decreases (higher up on the list), the language used in the news to discuss a company becomes more negative, and vice versa. And the higher up the list (lower ranks), the language used to discuss a company becomes more subjective as well.

For profit, the most significant coefficient was a positive coefficient for news subjectivity, adding more credence to the **positive relationship between news article subjective language usage and "success" (in the form of lower rank as seen above, or increased profit as seen here).**

The same trend follows for market cap- more "valuable" public companies are discussed with more subjectivity in news articles, and this is again the most important feature.

Overall, we believe that we can say regression using the OLS algorithm "fit the data" somewhat consistently with the correlation process above, as the coefficients represent the same magnitudes and directionality that many of the non-negligible correlations from above had. Additionally, most of the MAE values were relatively small compared to the size of the values that were being predicted, although it is worth noting that none of the MAE values were negligibly small, indicating that the regression models were not fantastic at fitting the data, but adequate.

Let's briefly stop and take note of some of the emerging insights from the correlation analysis and the regression coefficients (note that we will decide whether we think these insights are by chance or meaningful later):

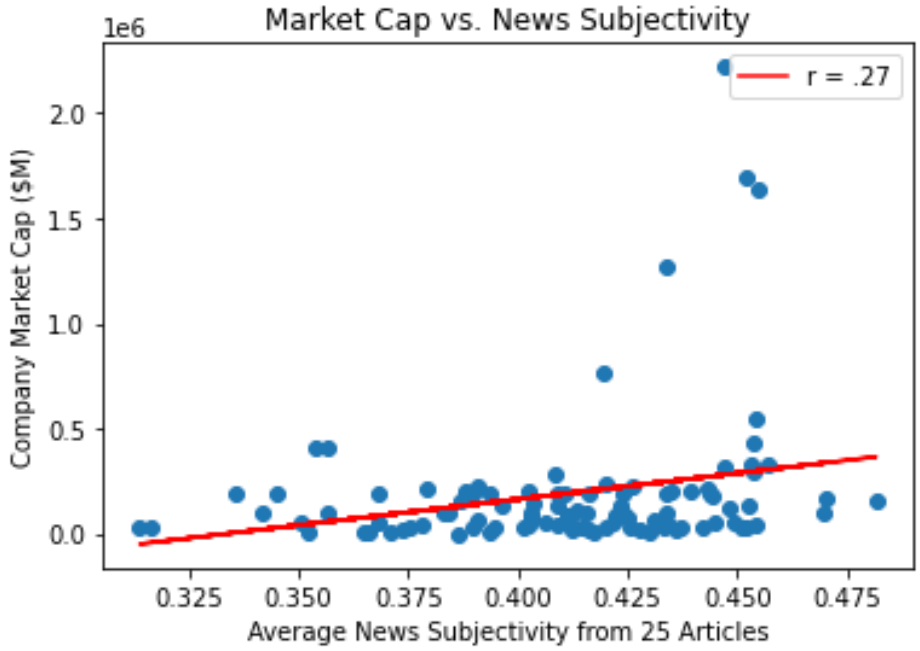
- The "is retail" dummy variable and polarity for both news and tweets seem to have a noticeable direct relationship. This suggests that retail companies do better in terms of sentiment polarity in tweets and in the news than the other sectors.

- The “is energy” dummy variable seems to have a negative relationship with news polarity. This suggests that companies in the energy sector tend to be discussed with more negative language than other sectors in news articles.
- The “ceo male” variable seems to have a negative relationship with news polarity as well, suggesting that companies with male CEOs tend to be discussed with more negative language than companies with female CEOs.
- Both the “is financial” and “is transportation” dummy variables have negative relationships with the news subjectivity score, indicating that these industries are discussed in more objective language in news articles than other industries.
- The “is technology” variable appears to have a negative relationship with media sentiment difference, suggesting that technology companies are spoken about with more positive language in news articles than in tweets.
- Variables indicative of a company’s “success” like low rank, profit, and market cap are directly related to news subjectivity, where the more successful a company is, the more subjective the language used to describe them in the news is.

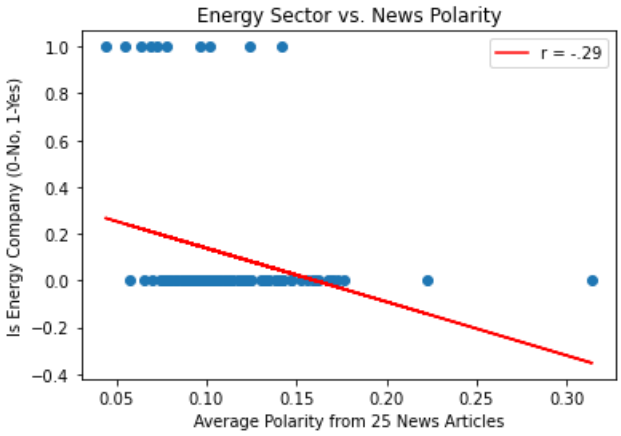
Insights Visualized

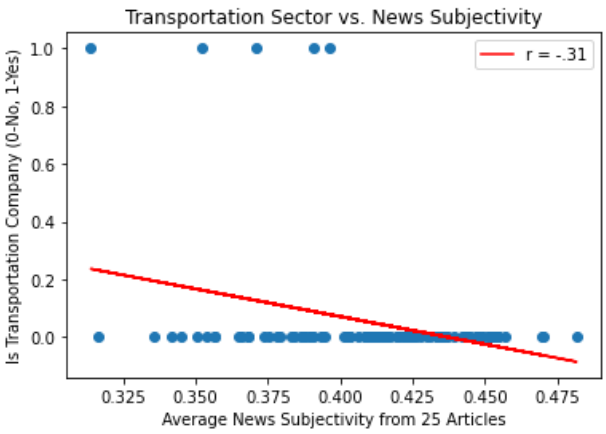
Let’s try to visualize some of the most significant findings thus far. Hopefully, we can get a sense of how our models are performing and what some of these relationships look like.

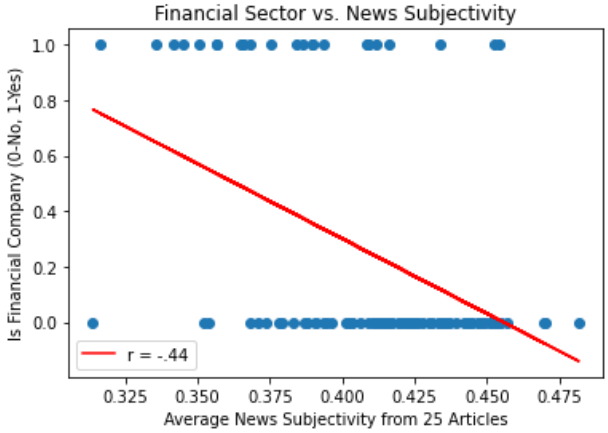
To represent the “success” vs news subjectivity relationship, this is a scatter plot of market cap vs. news subjectivity score:

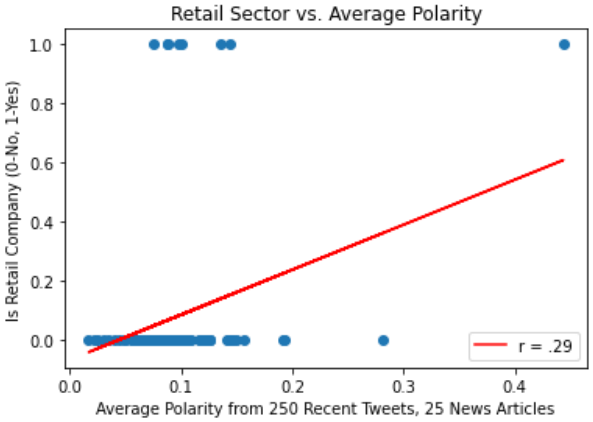


Next, we'll look at a few of the sector-specific relationships we have found.

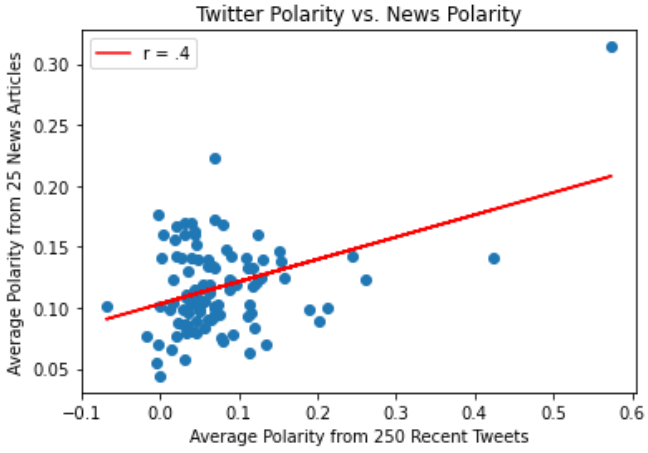








Finally, we'll look at the relationship between polarity in tweets and polarity in the news:



The features seen in the above plots characterize the most useful features in our investigation so far. All of the sentiment features are technically "engineered," so we can say that our engineered features performed well, as things like polarities and subjectivities have made it through correlation and regression and proved important. However, the more explicitly engineered features like media sentiment difference and intensity, and some of the averages between tweets and news, aren't super useful, and either don't indicate any relationships/signal, weaker relationships, or merely show what some of the simpler features already show. We will proceed (mostly, MSD may be worth checking for the Technology sector) with these simpler polarity and subjectivity scores in the rest of the report.

A Few Deeper Dives

As most of the algorithmic heavy lifting has been completed at this point, with the correlation process giving us initial direction, and regression analysis supporting some of our findings, we felt like a logical next step would be to begin the process of "diving deeper" into specific industries on a limited scale, as this sort of targeted research would be best suited for data science teams utilizing our process in an industry setting with more resources and data. We will only continue to validate the process we are developing, carrying out the natural next and final steps, as anyone in an industry setting looking to provide valuable insights for their business would.

To gain an initial understanding of how different industries compare against one another in terms of our sentiment variables, we can group by sector, average over the groups, sort by average polarity, and see the following:

sector	avg_twit_polarity	avg_news_polarity	avg_polarity	avg_twit_subjectivity	avg_news_subjectivity
Food, Beverages & Tobacco	0.035679	0.099571	0.067625	0.191474	0.421594
Transportation	0.060922	0.087170	0.074046	0.223549	0.364991
Energy	0.076959	0.084724	0.080842	0.276212	0.421963
Technology	0.042075	0.132846	0.087461	0.198164	0.433560
Aerospace & Defense	0.063575	0.120744	0.092160	0.239310	0.414409
Health Care	0.080927	0.109378	0.095152	0.268204	0.421184
Financials	0.073062	0.124152	0.098607	0.231568	0.384047
Telecommunications	0.095632	0.119495	0.107563	0.330133	0.414438
Food & Drug Stores	0.118306	0.119899	0.119102	0.309533	0.417056
Retailing	0.145450	0.147528	0.146489	0.288305	0.428290

Like we've uncovered before, we see Retail with polarity scores significantly higher than the other industries, and Energy at the lower ends. Food & Beverages was not consistently significant in the correlations and regression, so we will focus on the high overall polarities of the retail sector, and the low news polarity of the energy sector. We also can recall the negative media sentiment difference

correlation for technology companies, so we will look at that as well. Finally, the largest correlation we uncovered in our entire investigation was that of financial companies and news subjectivity, so we will explore that too. Let's take some brief deeper dives into these industries, representing the beginning steps of a further investigation following the process of "meta-analysis" of online sentiment data that we have developed here.

Unpacking Tweets and Keywords: A Closer Look!

Remember those min and max tweet and news sentiment features we engineered at the beginning of this whole process, and have been carrying along this whole time? Now that we are taking a deeper look into some specific industries, they will come in handy to potentially provide more context as to why certain trends are emerging.

Let's take a look at a WordCloud of keywords across the retail companies:

And some positive tweets:

- rt @rishaaaaaaa_: streaming on spotify and amazon is very important! \nplease share the links but not a snippet of the song. not even 5 seco...
- rt @jtimbre: 💰 my weekend flash #giveaway starts now! retweet, like, + comment to enter to win 1 of 12 @amazon $50 gift cards!!!
- @melanieszymans1 home depot just smells so good 😩
- rt @hardikpandya7: 🎯 on target with that win 😎
- @goosewaynetv welcome to lowe's! 😂

It appears that a lot of the sentiment behind the retail sector's discussion, at least in the news, is the inherent marketing and PR campaigns built into every little aspect of the language surrounding these companies. From the "best" in Best Buy or "target" of Target, to the catchy sloganeering in "Welcome to Lowe's!" everything is designed to bring people into physical locations, purchase

products, and direct consumer energy in a positive way unique to such an interpersonal industry.

Now let's take a look at a WordCloud of the Energy sector keywords:

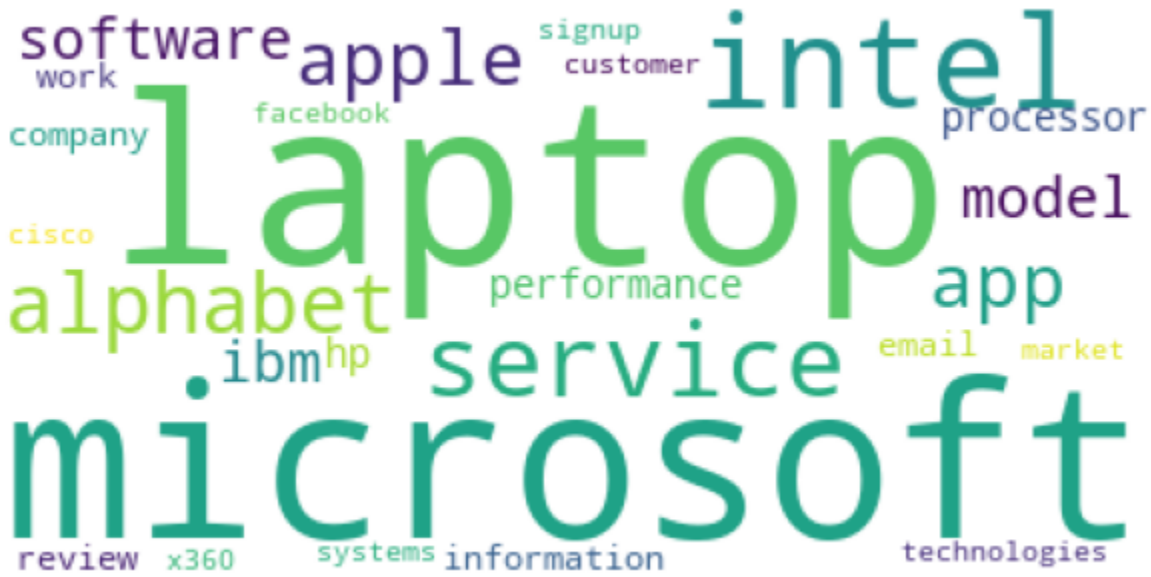
Energy companies appeared to suffer in terms of news polarity, and looking at some of the keywords being discussed and contributing to that negative polarity, we can begin to understand why. Companies, like ones in the Energy sector, that command widespread utilities while also being tied heavily to public markets, seem to be under a level of institutional scrutiny that other companies are not. Whether news articles are reporting about the suffering marketplace, and how energy company stocks are falling or traders are moving less of their volume, or whether it is merely pushback against outdated energy practices and discussion of a topic that is highly political, the language in the news seems to be noticeably more negative.

Now to dive into the media sentiment difference of technology companies, we'll compare some tweets with the news keyword WordCloud:

Negative Tweets:

- @microsoftedge @microsoft if its anything like windows 10 ill give this malfunctioning piece of crap a skip.
- rt @mattyharby: @aaronjmate this is absolutely laughable. every fbi and cia intel operative was diligently looking for collusion for 4+ yea...
- rt @manablazemick: @stefk88684429 @paratus2014 eskom fails to pay oracle but finds it "regrettable the manner in which oracle handled the m...

News WordCloud:



Taking a look at the tweets, we may come to realize that technology companies, likely more so than any other industry, are discussed intentionally negatively on social media like Twitter for a couple potential reasons. Users of social media tend to be users of technology, and these users will probably voice their negative opinions of technology products and services on social media more vocally than their praise, as, in our experience anyway, negative sentiment is a lot more likely to drive some sort of action or voicing of opinion. It doesn't help the technology company social media polarity score that oftentimes these companies will have established support accounts for users to tweet at with their tech problems in hopes of resolution. The first tweet and news keywords lend credence to the idea that technology products and services are consumed in a unique way, with certain levels of satisfaction with functionality and condition that might not be present in other industries, as technology is so crucial to so many of our daily tasks and processes, especially during the COVID era. In the news, articles discussing new products and the ever-advancing endeavors of technology companies naturally come with more positive sentiments, as we see things like "laptop", "software", "service", "performance", we can see that the article discussions are highly product and performance driven- there are no angry home-users tweeting at Apple to help them power on their MacBooks.

Finally, we'll look at some of the news article titles for financial services companies:

- 'Goldman Sachs Group Inc (GS) Q1 2021 Earnings Call Transcript - Motley Fool
- American International Group (AIG) gains 0.9% for April 07 - Equities.com
- Should You Consider Investing in Warren Buffett's Berkshire Hathaway (BRK-A)?- Yahoo Finance
- Is Wells Fargo and Company (WFC) A Good Investment Choice? - Yahoo Finance

- Colliers Mortgage Closes $19.9 Million in Fannie Mae Loan for The Ridge Apartment Community in Goodlettsville, Tennessee - MultifamilyBiz.com
- Morgan Stanley Posts $911M Loss Tied to Archegos - Bloomber
- Goldman, JPMorgan Chase stick it to the WallStreetBets crowd with knockout trading results - Fortune

We notice a potential source of the objectivity that we have analyzed in online news article discussion of financial companies. There seems to be a mix of instructive articles geared towards making certain opinions/practices more widespread, and structured reporting on various companies' moves in the stock market. These "educational" articles that are geared towards people looking to learn about an attractive industry to follow closely are written almost as more casual research papers would be, laying out the facts behind why one should invest here or there, sell what and when, etc. These reports are inherently more objective in nature to be persuasive, as the articles can better achieve their goal of being educational, informative, and even persuasive when little to no subjectivity exists in the article's discussion. In terms of the reports on various financial companies moves in the stock market, and their gains and losses, the stories here will naturally have less subjective content, as the reporting is usually a simple relaying of facts- percent changes, dollar amounts, corporate decisions, etc. These articles exist in such high volume for the people who follow the stock market closely and trade. They look for this objective information to inform trading strategies, which in turn fuels the production of more and more of this reporting on financial company successes and failures.

Overall, we have seen how deeper dives and investigation can be done for specific industries, or even individual companies (if there was enough data) in an industry setting that had more resources, computing power, data, time, etc. The above investigation using the tweet text and news article information shows the natural concluding step of the process we have developed. Once the models and algorithms give initial insights and potential relationships to explore, someone with specific company or industry data, specific sets of articles or tweets in certain timelines or by certain individuals, etc, can perform real and useful research for a company. We will summarize the specifics of the future research and investigation that our example deeper dives might inspire in the next section.

The End: What we did, and what's next.

We have reached the end of our data mining process, and it is time to reflect on the method for processing and mining insights from large amounts of data that we have developed, and also to discuss the potential insights that our process, even in its limited, academic capacity, are beginning to indicate, as well as the potential further implications of these insights, and of the overall process in general.

Discussion: Process, Insights, Extensions

The Process

In this project, we were able to work with data that many people might believe to be too messy, unstructured, unlabeled or overwhelming to generate useful insights.

In applying certain natural language processing (evaluating the meaning and intention of text-based data) algorithms, combining the outcome of these algorithms with more conventional datasets grounding the text data in real world industries, we moved this unstructured, messy data into a more manageable data "space", resembling the classic rows and column structure. We could then apply some simple, more conventional data analysis models/algorithms to draw out potential insights and point to various relationships that warrant further investigation. We then show one way to begin to carry out that investigation, namely by taking a look back at some of the important raw textual data that we had been carrying along with us the entire time. The process is highly replicable, to the point where we believe it could be carried out at large, efficient scale in an industry setting and in corporate research/data science departments.

The Insights

Although insights we generated during this process serve more as a validation of the process itself rather than findings that contain inherent value (which is an open discussion as to how realistic it is that truly valuable insights can be generated in the corporate space with purely academic resources), we feel that the trends we found are worth noting, and their significance/value is worth discussion. Here are a few of the most noteable ones that our findings suggest (the others are discussed throughout earlier sections):

- Retail companies are talked about positively on Twitter and in news articles compared to other sectors.
    - Other sectors (like Energy, for example) can look into the PR, Marketing, and Sloganeering strategies that retail companies employ, as they drive positive sentiment online.
    - Retail companies themselves can choose to either double down on their coordinated PR and Marketing efforts to further drive positive sentiment, or maybe decide to allocate resources elsewhere, as it is an aspect of business they are already outperforming in.
- Energy companies are talked about negatively in news articles compared to other sectors.
    - Energy companies might consider allocating resources to handle their negative discussion in the news, whether it be through a shifting of focus away from market losses, or promotion of positive sustainability efforts and projects, etc.
- Technology companies are talked about more positively in news articles than on Twitter.
    - Tech companies should further investigate what impacts this relationship between social media sentiment and news sentiment has with the success of their products and services. Despite us not finding any large relationships between online polarity and success indicators, on more individual scales for certain lines of products or different scales, what people see on social media may have real influences.
- Financial companies are discussed with much more objective language in news articles than other sectors.

    - Financial companies should investigate whether or not the sea of online direct reporting about markets, gains, losses, etc, and the relatively limited subjective pieces has an impact on their business. We were unable to link news subjectivity to indicators of success directly.
    - Depending on their investigation, actions could be taken to further promote this type of objective reporting, or stifle it through promoting more subjective, opinion pieces in the financial world that could potentially reshape the media sphere surrounding financial companies.
- Sentiment polarity (how positive and negative the language is) in the news and on Twitter are somewhat positively correlated.
    - This is more of a general finding that is less related to a deep dive into a specific industry or comparing sectors, but interesting regardless.
    - Further investigation into how sentiment in the news and on social media are alike, differ, and why, is worth investigating to better understand how public and institutional contributions into the "marketplace of ideas" and media relates to the corporate world.
    - It is worth noting that the two measures do not correlate with each other substantially, so investigation into why they differ may provide insights for Marketing, PR, and editorial teams across social media and news media industries.
- Some indicators of success were linked with subjectivity in the news, but we did not find a strong link between polarity and these success metrics.
    - Some companies may reconsider their PR focuses/strategies or consider reallocating resources if further investigation into the lack of dependency on polarity and success validates the null result here.
    - This also motivates investigations of what sort of customer/user/media entity behavior does in fact influence, or respond to success metrics. Surely there are some relationships, even if they don't lie where we've explored.

The Legitimacy of our Insights: Did it work?

Data snooping is defined as "statistical inference that the researcher decides to perform after looking at the data (as contrasted with pre-planned inference, which the researcher plans before looking at the data)." While we would say that there is the possibility that some data snooping existed in this investigation (as any real-world investigation would have), we don't believe it affected the insights produced in the end. Just as much as we picked and chose certain correlations, variables, trends, etc, to move forward with and continue discussing here, we also threw out initial leads that looked promising that led to nothing. We didn't force our engineered features to be the end-all be-all of corporate research, and in fact, we would say that the more explicitly engineered ones that we convinced ourselves were clever (like media sentiment difference, media sentiment intensity, etc) were not all that useful in the end, and the simpler engineered ones like the raw polarity and subjectivity averages were the most insightful. We abandoned looking at the promising leads with these features as those leads began to thin out, so we do not believe data snooping significantly impacted the insights produced in the end, and we believe that this process does not lend itself to data snooping overall, as the algorithms in use are

highly mechanical, and further investigation is up to the domain of the team performing the research in the end. Choosing to move forward with investigations in different sectors may be a relevant aspect of whatever corporate research is happening at whatever corporate PR, marketing, or data science department. We believe we showed how to adequately begin those steps, even if some data-snooping manifested in what domains and leads we kept emphasizing throughout, after seeing some initial promising correlations.

With the relatively small scale of our investigation in mind, we should make a decision about whether we think the insights we found are down to chance, or are actually meaningful. We believe that some of the stronger correlations and higher coefficient regression results, as well as the more sensible (at the expense of novel) findings do indicate real trends in the data. For example, retail's more positive sentiments, versus energy's negative sentiments in the news media, and the objective reporting of financial companies, as well as the weak correlation between polarity in the news and on Twitter. We believe these findings are indicative of some real (maybe not too novel or exciting) trends. For example, a report by TheAsci finds that major retail companies, despite having customer satisfaction scores hurt by the pandemic, posted very high customer satisfaction index scores across the industry for 2020-2021 [1]. Additionally, an executive summary of the Atlantic Council's report on the state of energy companies in the global transition effort to renewable energy (after they reviewed literature from articles and reports) states that "As the third decade of 21st century begins, the oil and gas industry faces opposition from a public greatly concerned with the environmental impact of fossil fuels, ever-more skeptical shareholders, and challenges from policy makers seeking to simultaneously meet decarbonization goals and expected oil and gas demand. Amidst a global energy transition, the demand, financial, and social future of oil and gas companies is increasingly in question." [2] While we don't have a direct source on the objectivity of reporting on financial companies, as we discussed before, it makes sense that so many earnings and losses reports, as well as educational articles use very objective language to provide very blunt information. And the link between polarity in the news versus polarity in tweets has very little online research to support it, but it had one of the highest correlations of the entire investigation, and makes logical sense. However, this can never be enough to accept a finding as true on its own, so we'll wait for more research before making claims on that relationship.

However, findings like technology company sentiment suffering more in tweets versus the news, some markers for company success being linked to news subjectivity, and the findings that we can't justify as much, were somewhat weaker, and make a little less sense overall, are most likely down to chance, with the right combinations of news articles and tweets being sampled to point in some arbitrary direction, or the distribution of either tweet or news sources being spread better or worse our relatively limited queries and investigation. Overall, there is both meaning in our findings, and likely some randomness, and we believe the randomness could be cut down on in more professional replications of a process like ours.

We believe that something nice about our process is that it would be easily adaptable to different data. While the findings themselves might change (in a different time period, other populations, different search queries other than companies, etc) the process merely involves scraping tweets and news articles as they are produced, easily assembling the "datasets" from the vast online word of textual

data. It's tough to "quantitatively" assert this robustness, other than that we were able to produce non-trivial correlations above certain thresholds (>.2,.3,.4) while assembling data from a variety of sources- a process that could be replicated with other data and queries. In fact, as a validation of this sort of process, we see NLP and sentiment analysis methods are beginning to be used more and more in the real corporate world. For example, take a company that Noah actually applied to for an internship, *LiquidNet*, an AI-based fintech company that provides asset managers with NLP based technology to scrape company documents, news articles, online threads, and social media to better predict the market [3].

Even with a robust process however, too high uncertainty will always plague every step of the process. If our queries are insufficient or incomplete, the articles or tweets returned don't reflect the population well, or the third-party joined data (like the Fortune 1000 data) isn't reliable, the process can't tell us anything. Textual data is inherently hard to process and unreliable, so when the third party data, querying structure, etc, also cannot be trusted to effectively elevate and unearth insights from the text, the entire process is hindered, no matter the expertise running the project or the resources at hand.

1. https://www.theacsi.org/news-and-resources/customer-satisfaction-reports/reports-2021/acsi-retail-and-consumer-shipping-report-2020-2021/acsi-retail-and-consumer-shipping-report-2020-2021-download
2. https://www.atlanticcouncil.org/in-depth-research-reports/report/the-role-of-oil-and-gas-companies-in-the-energy-transition/
3. https://www.liquidnet.com/

What's Next, and for Who?

The logical continuation of our work here is of course, an application of our methods to an industry setting with more resources, computing power, storage capabilities, better data, and more time. We have been plagued throughout with certain difficulties that likely often befall academics engaging in corporate research processes. Corporate data like the information that would be extremely useful and interesting for a project like this one (info on charitable donations, political involvement, lobbying history, etc) that could really show some interesting relationships with sentiment data and measures of success are simply unavailable to us. They're either behind expensive paywalls, restrictive APIs, or simply under lock and key never to be released to the public, and unfortunately, this is the nature of the data in this area of research. When the spirit of research takes a back seat to the overarching goal of being profitable, and running a business, lucrative data and information is hard to come by. However, there are methods to gain access to these things with the right resources and in the right settings.

The optimal audience for our method/process/project are the PR, marketing, and Data Science departments at successful corporations with large budgets for data and research, like those in the Fortune 1000. We would definitely characterize the volume we have created (in terms of the process we developed) falls under the category of decreased uncertainty (a process to unearth insights from extensive textual data) and saved time (we assembled an algorithmic process to sort and examine potential large sums of data efficiently). The process we have developed

here, if not already in practice at these institutions (which we're sure some version of is definitely at most businesses) could prove an incredibly valuable method for extracting the ways that measures of success, and extraneous-joinable variables like donations, genders, etc, relate to sentiment information. Natural language processing continues to emerge as a field of data science being implemented in the research departments of large corporations, and for good reason. There is an endless supply of text data floating around on the internet full of insights to be mined.

In terms of the actual insights that we mined as a sort of validation of the process we developed, the audiences are of course the respective industries in question, specifically PR and marketing teams at successful retail companies like Amazon, Target, Best Buy, Home Depot, etc, as well as the captains of the energy and technology industries, whose names we are sure everyone is familiar with. The value of the actual insights we showed as validation fall under the category of decreased uncertainty, as these insights may educate members of these industries about some sector-wide sentiment specifics and trends they previously may not have known existed. Of course, the actual insights present in the report are not as useful as what in-depth research by the ever-more powerful data science teams in these companies could produce with the method we have demonstrated here.

Some Refinements to the Process

Discussion of next steps wouldn't be complete without some talk about what could be improved with this process if we had more time and resources (what we would have done differently).

Our investigation could be carried out on data pre-COVID (our first iteration was in fact with 2018 Fortune 1000 data on accident), and compared to a during/post-COVID corporate and media landscape. We could then potentially see period-specific shifts, and come to better understand how world events, or merely passing time, affects certain sentiment, success, and sector relationships.

Additionally, someone iterating on our process here could consider weighing tweets by the amount of likes, retweets, and replies they have to better represent influence. The same weighting could go for accounts with more followers, or news sources with more average daily readers, etc. This could come with the generation of new "popularity" or "influence" related features.

Conclusion

Overall, in this project, we were able to piece together a process for analyzing large amounts of text from the internet, and drawing out potential insights related to the way different businesses are discussed in different forms of media. We pulled tweets and news articles from the web, looked at some data about successful companies, and together, tried to unearth some trends that tell us a little more about the subtleties between different industries, success, and forms of media. We were somewhat successful to this end, as some of the insights we compiled about industries and their online discussion made sense and had backing, while others were harder to justify, and some that we expected to materialize never did. We weren't too distraught at the prospect that our insights weren't revolutionary though, because we feel that the real value of our project

comes from the process we demonstrated. We showed taking text data from different sources, joining it with domain specific information about businesses (or whatever data you're interested in), and plugging the data into some algorithms to extract relationships and insights is a viable tool for data mining and analysis that would likely thrive in industry settings.

We are proud to have been able to use the things we learned in this class, along with our own natural curiosities about generating insights from data to work on a meaningful and applied project. This has allowed us to gain invaluable hands-on experience, and experience an application of some of the most crucial concepts that we had the privilege of learning in STAT 3106 with Wayne this semester. We are excited for whatever further applications of these skills and curiosities lay ahead for us.

Another Group's Project Critique: Isaac Horwitz and Begum Babur

Isaac and Begum examined the frequency of tweets demonstrating depression and anxiety, focusing on the person's gender, U.S. state location, and word usage. Their project is very important and interesting; there are few limitations that we noted. First, the data set is very narrow. Given that the tweets are only pulled from the past week or so, it may not reflect how people are generally feeling since that tends to fluctuate. Additionally, the current coronavirus pandemic has certainly increased levels of anxiety and depression, and that may be reflected in people's tweets. Perhaps an analysis and comparison of pre-covid and post-covid tweets would be interesting to perform, in order to account for the pandemic's notable effects on our mental health. Second, the method used to predict gender is somewhat flawed. There are many usernames on Twitter that are not real names and do not indicate gender. They relied on the model's prediction for gender, which can't be very accurate for those numerous usernames that do not clearly specify gender. This may affect the results and analysis. Finally, the project provides only a snapshot of thoughts people may be having, and they note that age is an important factor to consider. Twitter is not used by all age groups equally, and younger people and more likely to use it than older people. As a result, older people may not be well represented in this data, and further research and data collection are required in order to ensure that all age groups are represented. Overall, this project is informative, relevant, and an important investigation.