Marketing sentiment of Powertrain, Through NLP

Team Emotion: Jianglun Xie, Tiancheng Wang, Zhonghao Yang, Wanting Zhao, Huilin Ma

Abstract

Donaldson Company, Inc. is a vertically integrated filtration company engaged in the production and marketing of air filters used in a variety of industry sectors, including commercial/industrial, aerospace, chemical, alternative energy, and pharmaceuticals. Our goal is to understand the market sentiments towards alternative powertrains (battery vehicles, fuel cells, hydrogen combustion engines, etc.). We had three technical parts including, data extraction, data cleaning and sentiment analysis to help Donalson adapt its product portfolio and strategy towards electrification; leverage big data to create better insights across the entire powertrain industry.

Executive Summary

What is Sentiment Analysis?

Companies everywhere are seeking to leverage the power of Machine Learning, especially opinion mining. Obviously, they seek to gain insight from reviews, comments, feedback surveys, and other textual data sources to work both more efficiently and more effectively. It can also open new avenues and provide the organization with an additional competitive edge.

With the new business models for alternative powertrains coming, Donaldson wants to understand truck and heavy equipment company sentiment for alternative powertrains. We will collect data from several different social media, we are focused on Twitter, using open-source tools to establish an automated data collection and synthesis process.

This process is quite challenged and there are many issues to overcome. The technical challenges of developing platforms and formulating solutions, and the challenges of implementation and management improvement. and the sentiment analysis section.

Sentiment analysis is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to the voice of the customer materials such as reviews and survey responses, online and social media e.g., news text where authors typically express their opinion or mood.

The key to data mining —API – application programming interface

APIs enable your product or service to communicate with other products and services without having to know how they're implemented with agreements between parties. (**Figure 1.1**)

Natural Language Processing(NLP)

This is where natural language processing(NLP) comes in. NLP allows computers to understand text and spoken words in much the same way human beings can. We used NLP to analyze the text data that we collected from Twitter. After using the NLTK Sentiment Analysis Package to evaluate the sentiment compound values of the text data, our team segmented the text data into Negative, Neutral, and Positive. In order to visualize the data, our team used the Matplotlib, Plotly, and seaborn Package in Python, in addition to Tableau.

Powertrain Choice

By using the APIs and different methods of NLP, we are able to explore more specific analyses to help Donaldson find internal and business outcomes of the powertrain. We are committed to finding out the explanation and logic behind the texts and numbers. Visualized the logic behind the number of declines or increases and the cause of people's attitude change from positive to natural or negative on powertrains.

Based on our exploration, battery electricity is the most popular on Twitter, and it could be a potential business development opportunity. So, we suggested Donaldson leverage battery electricity to develop the business for now.

Data collection and cleaning

Access of LinkedIn & Twitter

Due to the policy restriction of LinkedIn, the required data that is used to do sentiment analysis cannot be accessed from LinkedIn by using API. The target platform for obtaining data was switched from Linkedin to Twitter to avoid this problem. The data that included tweets and comments were successfully collected from Twitter by using API in Python, and the sentiment analysis for the powertrain usage in the future was conducted successfully based on the data that were collected from tweets of the target company's official account.

Data Extraction

There were several documents that needed to be prepared for this project before the process of sentiment analysis, which are the company list and, keyword list, the scope for the data extraction from Twitter by using API. The company list targets heavy equipment companies from all over the world, such as Volvo and Mercedes-Benz for the data extraction in this project. The list is used for the scope of data collection by scraping the related tweets and comments for the topics, which are related to heavy machinery equipment and powertrains of their Twitter account.

The keywords are several types of powertrains, such as hydrogen fuel, battery electric and natural gas. It was used in API to filter out unnecessary information from Twitter and make the collected data as accurate as possible. This step is aimed to conduct sentiment analysis to see whether people have positive, negative, and neutral sentiments towards these powertrains.

After having these two documents, the next step is to clean the company's name list. Since there are some account redundancy and missing in Twitter, the company's name list should be cleaned, including deleting the repeated company, unifying the company name with their official Twitter account, and deleting the company name that doesn't have the Twitter account, such as some Asian companies. This step could help to scrap the data accurately, quickly, and effectively by using API in Python.

Data Cleaning

Data cleaning is the process of re-examining and verifying data. Its purpose is to delete duplicate information and modify existing errors, which aims to make the dataset accurate and clear to conduct sentiment analysis.

During the data cleaning process in this project, some columns were deleted or renamed, for example, author id to make the dataset clear and increase the model's accuracy. Then conduct data wrangling, involving dealing with missing values, lowercasing all text letters, removing hashtags and punctuations, etc, and output a final version of the cleaned dataset, which includes approximately 47,000 rows of data, and 3 columns, which are date, original text, and edited text as the database to conduct the sentiment analysis and modeling in the next step.

Technical Deep Explore

Social media is one of the most consequential platforms that contain various unstructured data information. In order to get helpful information from selected social media platforms, text analysis algorithms, natural language processing (NLP), and statistics insights were used to analyze customer and consumer sentiment by classifying text into positive, negative, or neutral categories.

Understanding consumer and customer reactions on an emotional level are critical for unearthing the most profound insights directed to adequate preparation for the industry trendsetter. Therefore, as the data set was finalized and transformed into the desired data structure after the procedure of text analysis, we implemented various text-mining methods to explore the deep meaning behind Twitter tweets.

Sentiment Distribution

Transforming the textual information into numerical information is the first step before the actual sentiment analysis; according to the NLTK Sentiment Analysis Package, we got the polarity score for each tweet. Then we expanded the polarity score to **Negative**, **Neutral**, **Positive**, and **Compound** values for further classification. In order to get accurate compound scores, we choose VADAR package in Python, it is a lexicon and rule-based sentiment analysis tool that is specifically for social media. VADER uses a list of lexical features, such as words, which are generally labeled according to their semantic orientation. It not only tells about the Positivity and Negativity score but also uses compound scores to tell us about how positive or negative a sentiment is. Finally, based on the output of the compound value, we classified the tweets as

Negative if they were between negative one and zero, **Neutral** if it equals zero, and **Positive** if between zero and positive one.

As we finished the basic structure of the sentiment analysis, we concluded the sentiment distribution in **Figure 3.1**; the sentiment of neutral and positive contains the preponderance of the whole data set, which is 45% and 44%, respectively. A potential reason for the percentages is that some alternative powertrain keywords are proper nouns containing neutral or slightly positive sentiments. For example, the top 2 highest frequency words are "electric" and "battery"; the relevant sentiment score is near zero, which will be highly possible to be classified as "neutral."

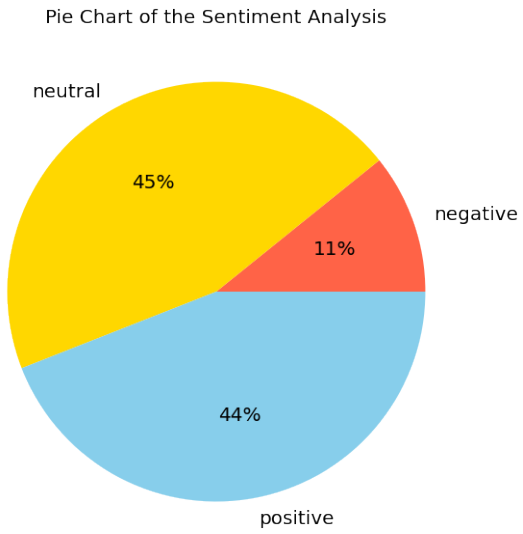


Figure 3.1

Machine Learning Model – Navies Bayes Classifiers

Before getting deep into our data and sentiment analysis, our team decided to build a machine learning model to satisfy the client's request for a prediction model and examine the accuracy of sentiment score we previously got. The first purpose of this step is that our team utilized

machine-learning strategies to identify an appropriate model to help predict the sentiment of future tweets relevant to the keywords provided by the client. In order to allow the client to effectively utilize the model in the future for prediction, our team recommended the Naïve Bayes Classifier. X-input and Y-output were interpreted as "the frequency of each unique word mentioned in the text" and "previously identified categories of the text, in this case, positive, negative, and neutral." The second and the major purpose of this step is using the result of the model to prove the sentiment score is trust-worthy. We used extracted data and split it into a training set and a testing set. After the training set well trained the model, the testing set will provide us an accuracy rate. If this rate has a high score, it not only means this model is well designed, but also means the sentiment score we previously got a high accuracy and trust worthy.

After implementing the data and experimenting with the Navies Bayes Classifier, our team received the confusion matrix of the model, as shown in **Figure 3.2**. The model of Navies Bayes Classifier is 77% accurate when predicting the sentiments of future tweets once the X is input correctly. This result is relevantly high enough to prove sentiment analysis provided us a accurate sentiment score to develop future analysis. The next step for our team would be to optimize the model to seek opportunities to improve the accuracy further to fit the client's needs.

	precision	recall	f1-score	support
0	0.50	0.61	0.55	2358
1	0.84	0.77	0.81	7903
2	0.79	0.80	0.80	8587
accuracy			0.77	18848
macro avg	0.71	0.73	0.72	18848
weighted avg	0.78	0.77	0.77	18848

Figure 3.2

Statistic Analysis Outcome

As we look at the boxplot of the sentiment distribution, **Figure 3.3** contains a boxplot that separately categorizes "Positive, Neutral, and Negative." We found that the "negative" boxplot contains many outliers below the third quartile boundary.

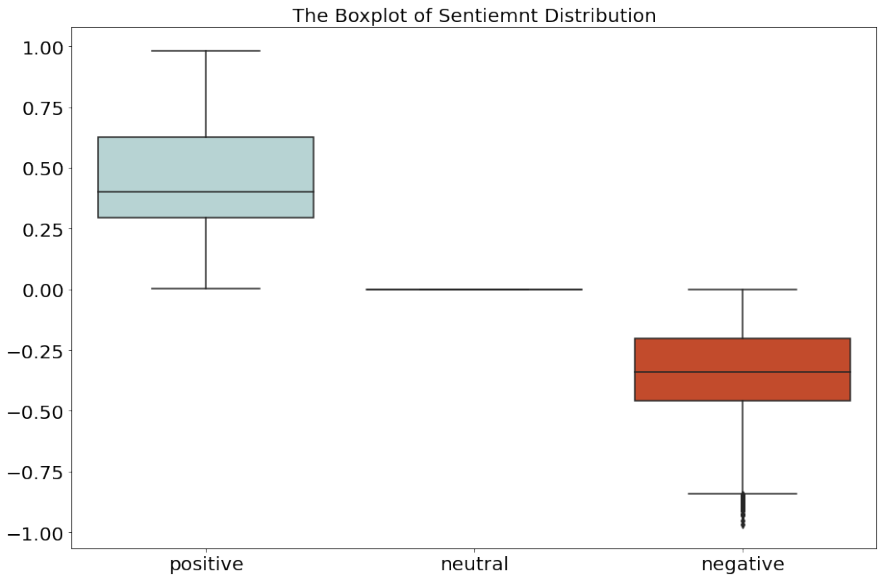


Figure 3.3

As our team took a closer look at the outliers, two tweets revealed that the underlying reason for the highly negative sentiment might be that alternative power trains, such as hydrogen and electricity, are still developing, causing security and sustainability concerns. The first tweet was published on 4/19/18, *"Some Chevy Bolts have battery problem that's every EV driver's worst nightmare: General Motors isn't calling it a recall, but Chevy Bolt electric vehicle owners are being notified about a battery problem that is affecting some of its cars on the road in";* the second tweet was published on 8/26/22, *"When I spoke with Volvo, that is what they told me, by 100,000 km, I would need to pay for a full replacement battery of $30,000. That,Äôs a new effing car! What a scam electric vehicles are and a waste on resources also too damaging to our earth if you look into it."*

In order to find out the sentiment changes between years, we build a timeline analysis based on each year's average sentiment scores. As we can see from **figure 3.4**, the average sentiment score is in the range of 0.1 to 0.2. Recall the maximum positive sentiment score is 1, the average sentiment score by years shows Twitter users has a slightly positive attitude towards powertrains. The graph also shows a trend that the sentiment score has a huge decline from 2019 to 2021, and a decline from 2021 till today. The possible reason for the decline will be analyzed in the business insight section.

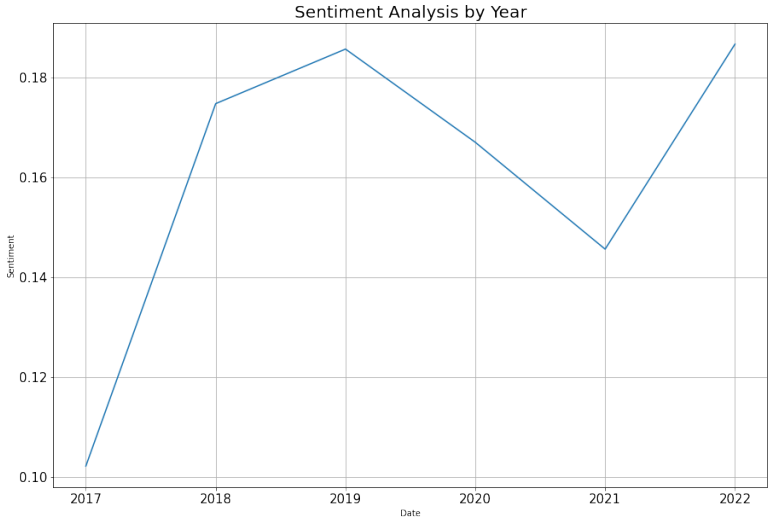


Figure 3.4

N-grams Distribution

Our team used the list of keywords provided by the client and scraped all relevant tweets between the year 2017 and the year 2020 to perform text analysis. Specifically, our team highlighted the most mentioned words among all tweets and found that the top 10 most mentioned words were similar to the keywords we initially used to identify the relevant tweets **(Figure 3.5)**.

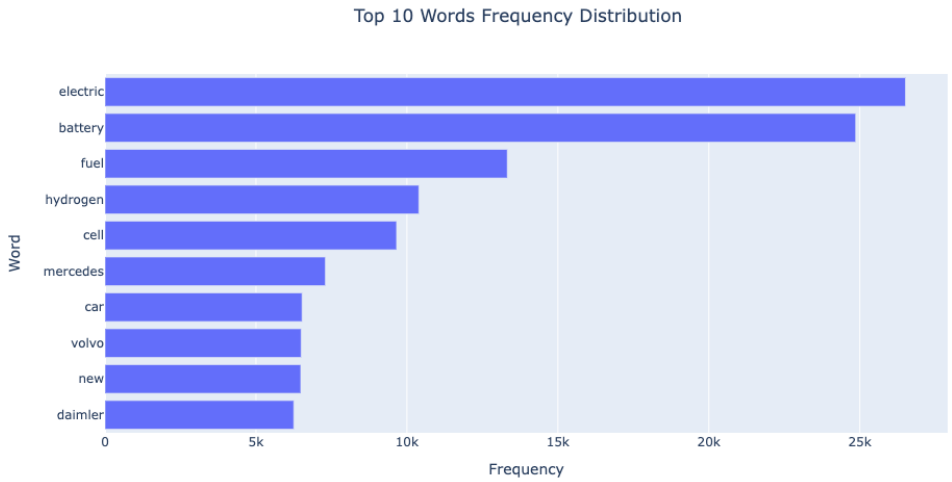


Figure 3.5

In order to dive deeper into analysis and gain more insightful information, our team executed the N-gram method to see if we could identify a different set of the most mentioned words among the tweets. As a result, our team found several brand names frequently mentioned in the tweets. For example, "Mercedes Benz" showed up in the top 20 two-word combinations (**Figure 3.6**); "Hyundai Motor Group" was included in the top 20 three-word combinations (**Figure 3.7**). Therefore, it is evident to our team to conclude that when people discuss keyword-related topics, they tend to include influential brand names in the discussions.

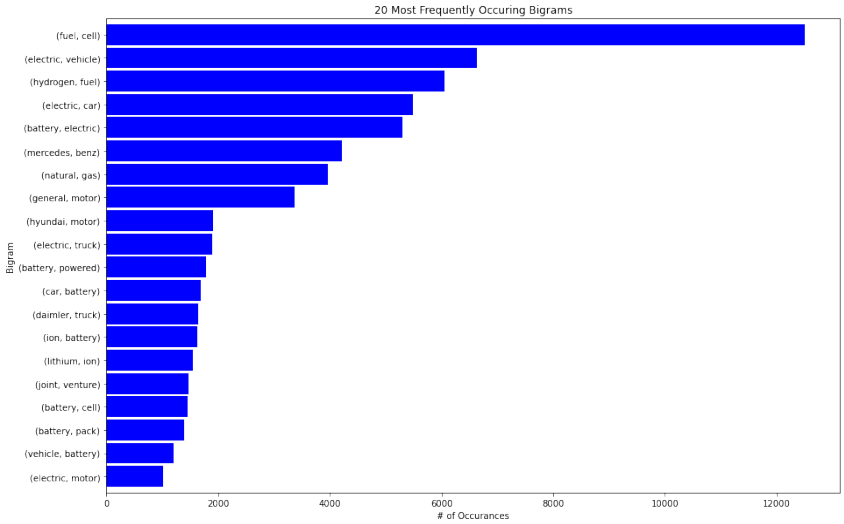


Figure 3.6

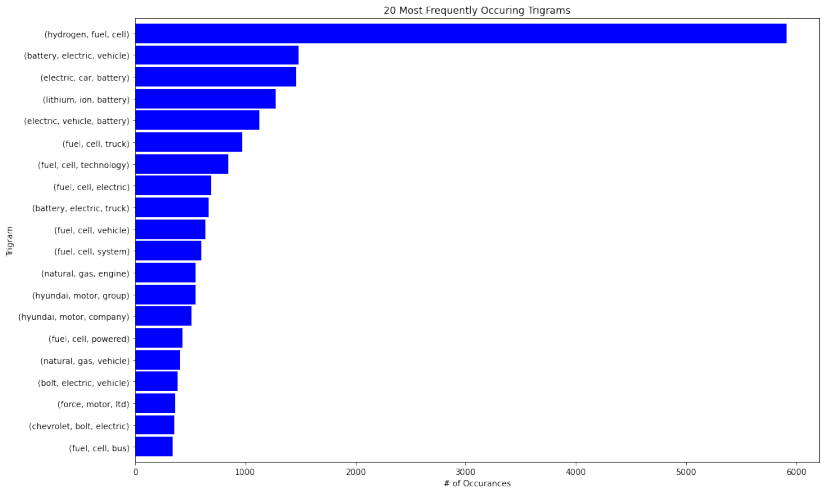


Figure 3.7

Sentiment Analysis By Segments

The original data was divided into 4 segments, based on different types of powertrains, which are battery electric, hydrogen fuel cell, hydrogen combustion engine, and natural gas.

Note: The “Electric” in the graph represents the battery electric, “Hydrogen” represents the hydrogen fuel cell, “Hydrogen Engine” represents the hydrogen combustion engine, “Natural Gas” represents the natural gas.

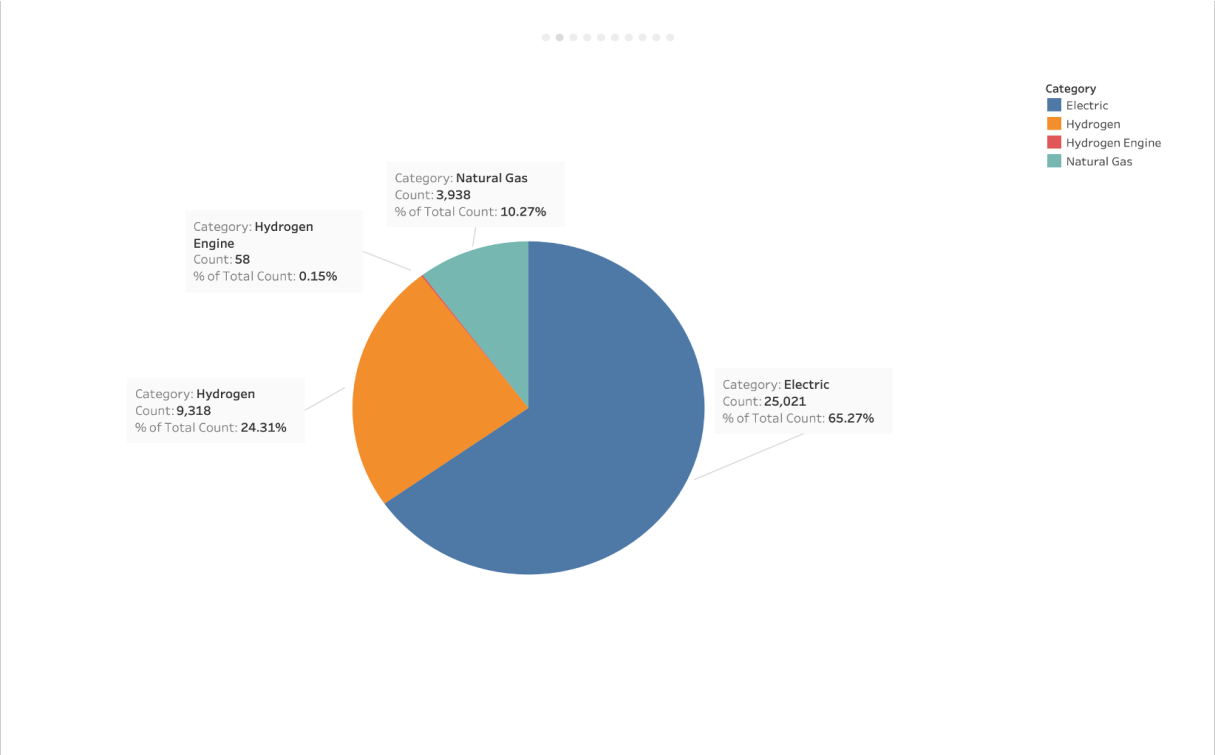


Figure 3.8

Figure 3.8 shows the proportions and counts of each powertrain segment. It was shown that the battery electric has the largest proportion here, which accounts for 65.27% of all segments. The second largest segment is hydrogen, which accounts for 24.31%. Size of battery eclectic data is approximately two thirds of the whole data set.

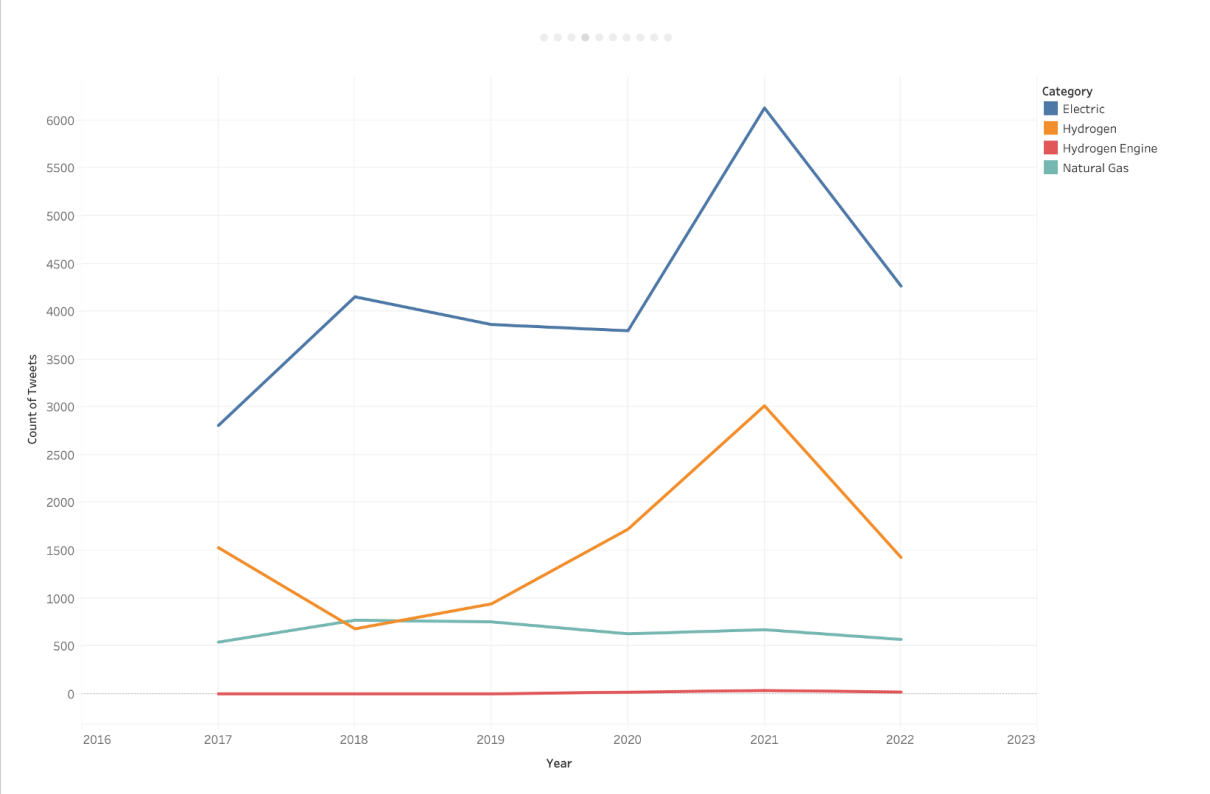


Figure 3.9

Figure 3.9 shows the count of tweets from 2017 to 2022. Battery electric and hydrogen fuel cells reached their peak in the year of 2021. Battery electric has the highest count of tweets among all four powertrains over the five years.

Both figures suggest that battery electric is mentioned more frequently and has a higher trend on Twitter. The comprehensive data visualization of the sentiment for battery electric is shown as below **(Figure 3.10)**:

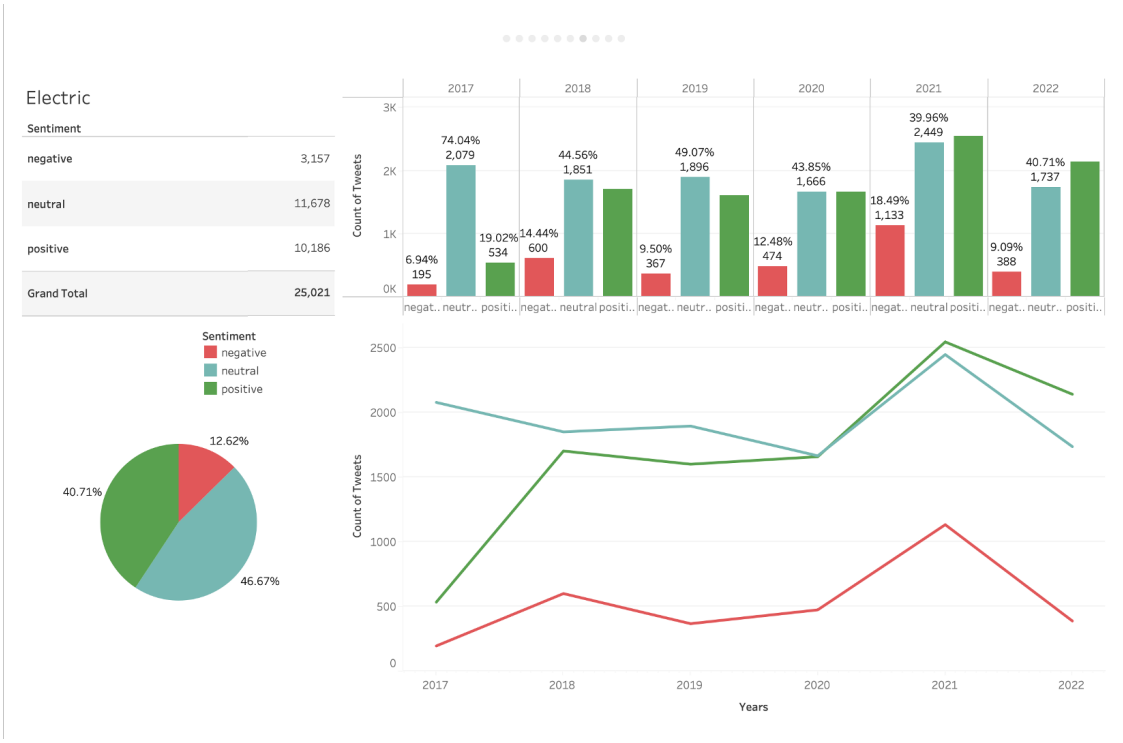

Electric
Sentiment
negative 3,157
neutral 11,678
positive 10,186
Grand Total 25,021
Sentiment
negative
neutral
positive
12.62%
40.71%
46.67%
2017
2018
2019
2020
2021
2022
Count of Tweets
3K
2K
1K
0K
6.94%
195
74.04%
2,079
19.02%
534
14.44%
600
44.56%
1,851
9.50%
367
49.07%
1,896
12.48%
474
43.85%
1,666
18.49%
1,133
39.96%
2,449
9.09%
388
40.71%
1,737
negat..
neutr..
positi..
neutral
2500
2000
1500
1000
500
0
Years

Figure 3.10

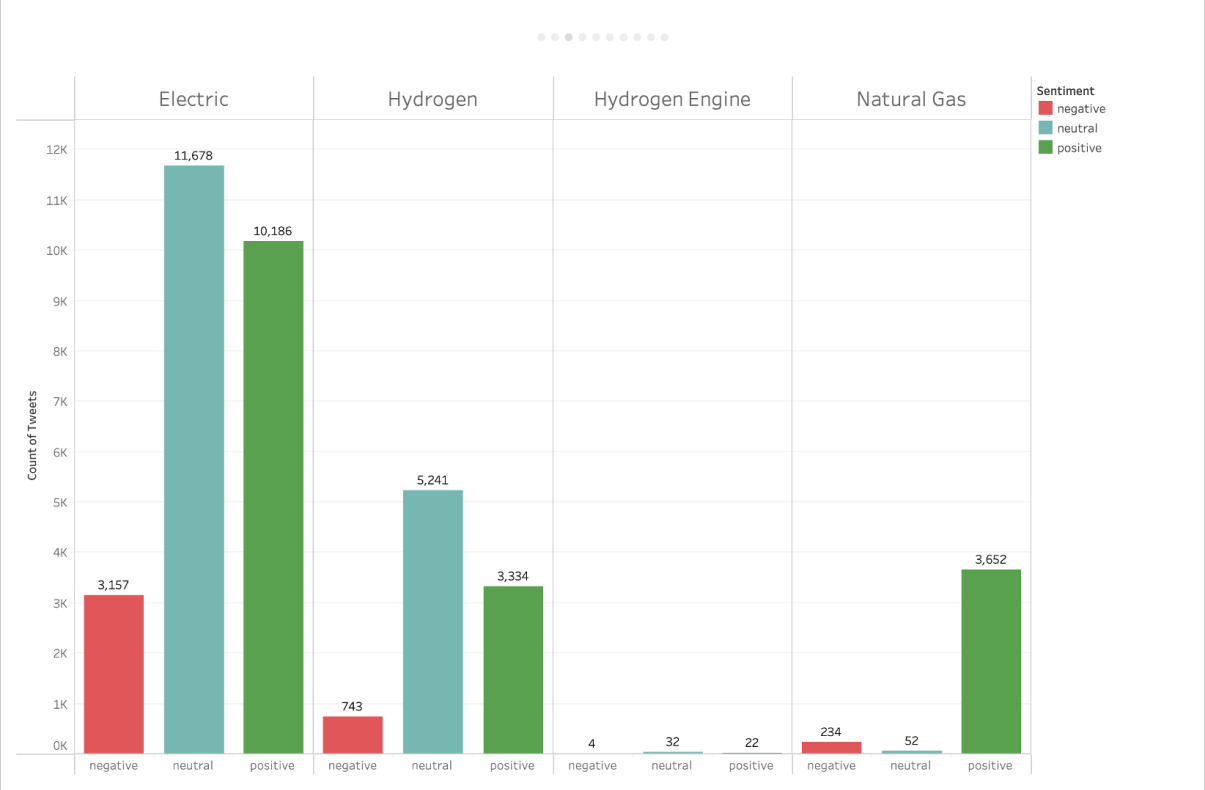


Figure 3.11

Figure 3.11 shows the sentiment counts for the 4 different powertrains. For the electric, neutral has the largest number among its segment. The second largest count is positive. For hydrogen, neutral also takes the largest proportion. And the second largest count of tweets is positive. For the Hydrogen Engine, there are only a few tweets about it. And among those tweets, tweets of neutral have the largest number. However natural gas is completely different from the other three segments, positive tweets take the largest number among all tweets towards natural gas. Overall, the positive rate of natural gas is the highest, which is about 92.7%.

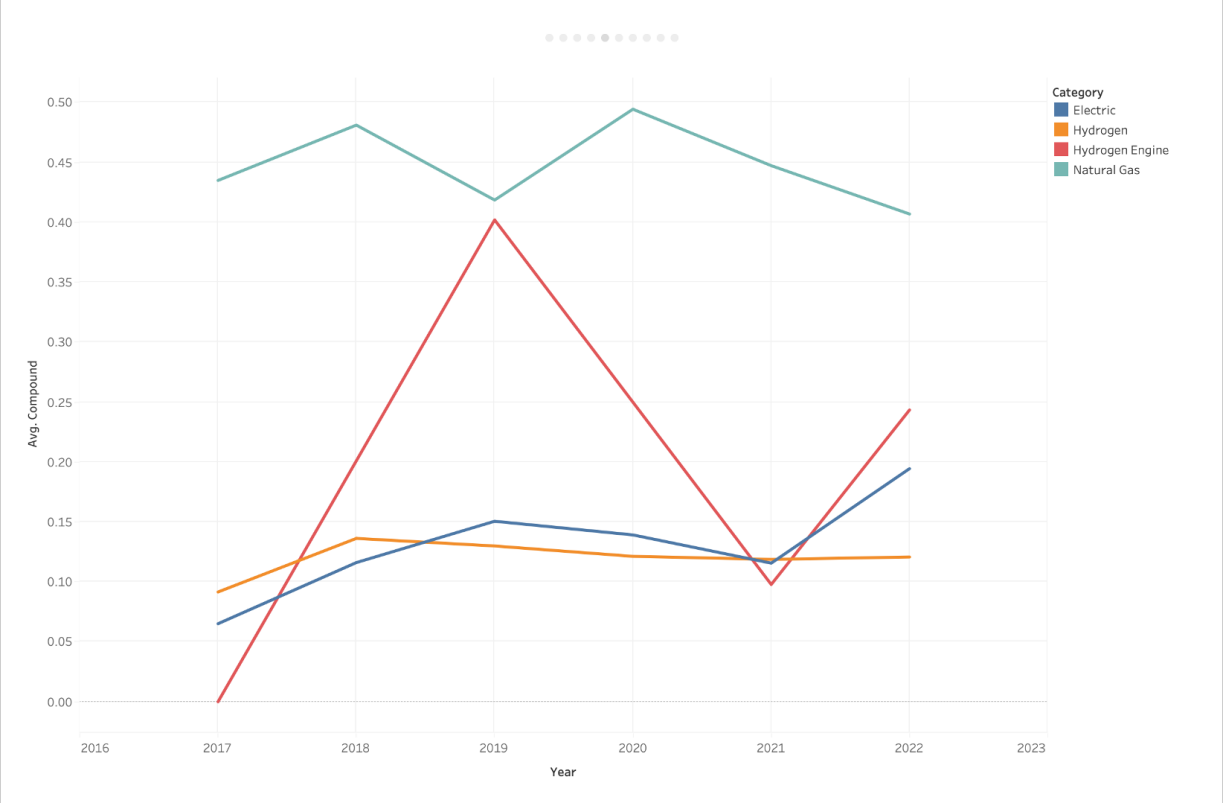


Figure 3.12

Figure 3.12 shows the average sentiment score of the 4 powertrains from 2017 to 2022. It's noticeable that natural gas has the highest average sentiment score among those 4 powertrains over the past 5 years. A closer look of natural gas's tableset is shown as follow **(Figure 3.13)**:

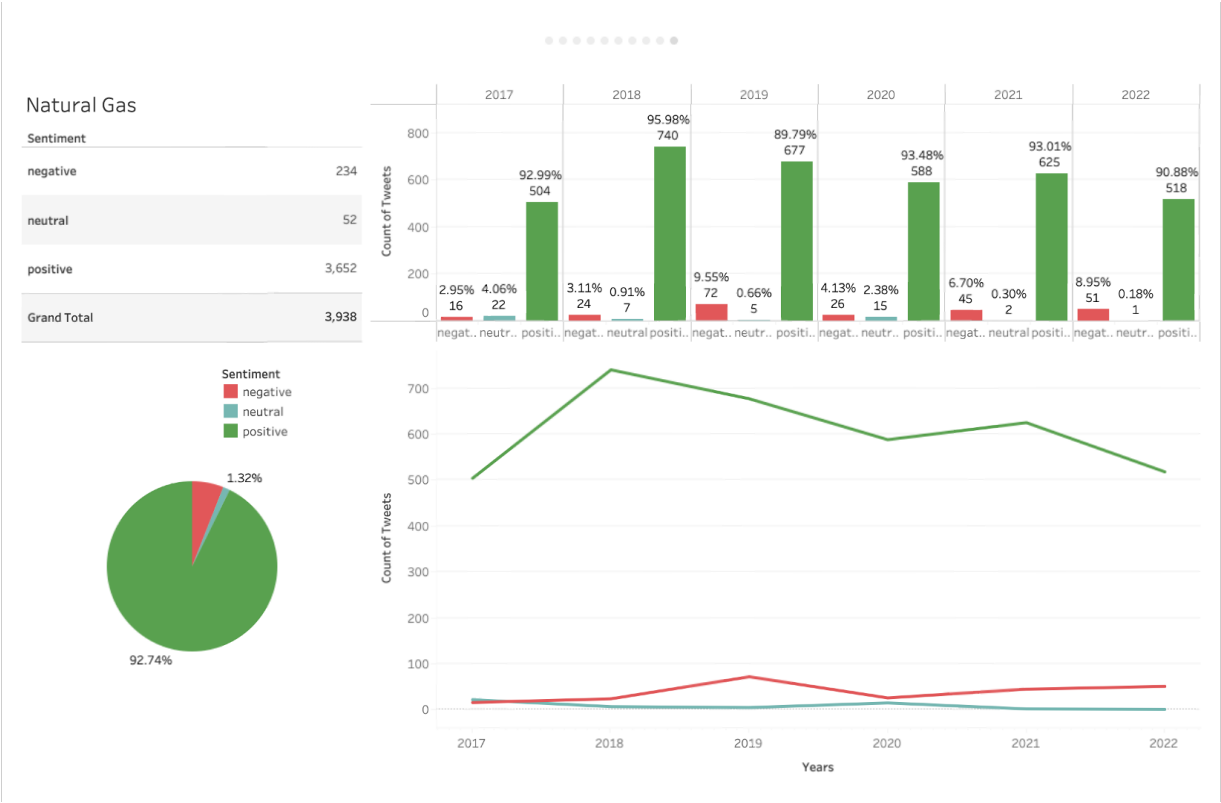

Natural Gas
Sentiment
negative 234
neutral 52
positive 3,652
Grand Total 3,938
Sentiment
negative
neutral
positive
1.32%
92.74%
2017
2018
2019
2020
2021
2022
Count of Tweets
2.95% 16
4.06% 22
92.99% 504
3.11% 24
0.91% 7
95.98% 740
9.55% 72
0.66% 5
89.79% 677
4.13% 26
2.38% 15
93.48% 588
6.70% 45
0.30% 2
93.01% 625
8.95% 51
0.18% 1
90.88% 518
Years

Figure 3.13

Conclusion and Insights

Combined with our result of different methods of NLP, we are able to make more specific analysis in order to find internal and business outcomes in this project. Our goal is digging deeper into our analysis results to find out the explanation and logic behind the texts and numbers.

Insights

Recall the timeline analysis **(Figure 3.4)**, the average sentiment score increased from 2017 to 2019, followed by a huge decline from 2019 to 2021, and an increase from 2021 till today. This trend gives us some confidence that the sentiment score is possible to increase through the rest of 2022 to 2023. The decreasing trend from 2019 to 2021 is highly overlapped with Covid-19 pandemic period, the pandemic is possible to be one of the major causes of the decline in sentiment score in these two years, directly and indirectly. In a recent-published research on *World Electric Vehicle Journal*, most major regions show a decline or growth freeze in semiconductor market share from 2019 **(Figure 4.1)**. Especially in industry electronics and automotive electronics, which these two segments are correlated with powertrains of trucks and heavy equipment . [1]The logic behind these declines is that the pandemic influenced suppliers and transportation of components and materials for powertrain components, such as components of electric vehicles. Those shortages are most likely to be one of the causes of people's attitude change from positive to more natural or negative on powertrains.

[1] Frieske, Benjamin, and Sylvia Stieler. "The 'Semiconductor Crisis' as a Result of the COVID-19 Pandemic and Impacts on the Automotive Industry and Its Supply Chains." World Electric Vehicle Journal, vol. 13, no. 10, 2022, p. 189., https://doi.org/10.3390/wevj13100189.

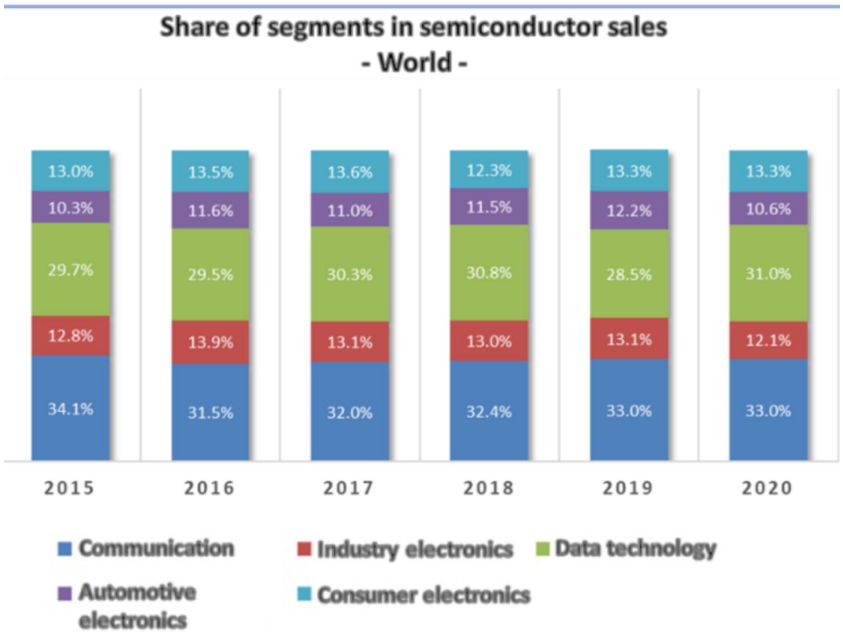


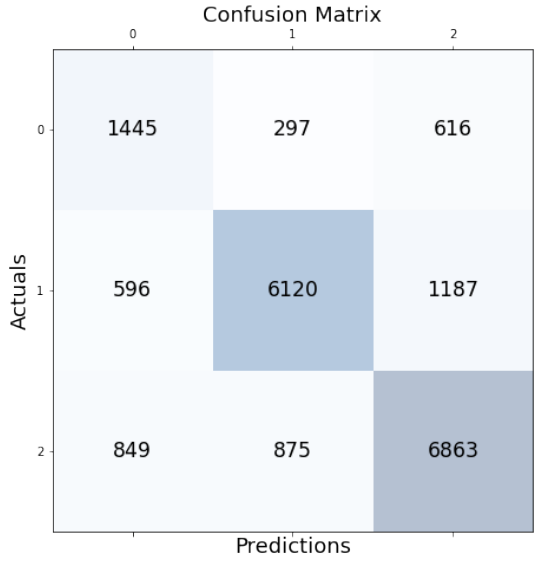
Figure 4.1

One important feature of the timeline analysis we must point out is that the average sentiment score by year is in a range of 0.1 to around 0.2, very slightly positive, which indicates most twitter users have a relatively natural and slightly positive attitude towards the industry of trucks, heavy equipment, and powertrain as a whole. This result also shows up in our Navies Bayes Classifiers model. A confusion matrix is made to test the accuracy of our model. (**Figure 4.2 & Figure 4.3**) Although the model has a 77% accuracy, we can still find this model has a high false-positive rate, which means some tweets supposed to be natural, but the model reported positive. This result indicates that a large portion of the tweets have a very slightly positive attitude, it's difficult for the model to decide their sentiment. Based on the results of the low positive sentiment score, we suggest that there’s huge potential to increase people’s awareness of the positiveness of truck and heavy equipment’s powertrain, which also means huge potential

market.

Figure 4.3

Figure 4.2



	Negative	Neutral	Positive
Negative	1445	297	616
Neutral	596	6120	1187
Positive	849	875	6863

Conclusion

Finding which powertrain we should focus on for further study is also one of our missions. In the previous section, we saw a comparison of top frequent twitter keywords in one, two, and three grams. Electric battery is mentioned more frequently. The sentiment analysis by segments proved 65.27% of the data is in the battery electric segment, much more frequent than other powertrains. It means battery electric has a much higher popularity than others, most of the tweets are neutral and positive. However, although natural gas has much lower frequency, it has a much higher average sentiment score. Twitter users have a very

positive attitude towards natural gas.

In conclusion, we suggest Donaldson focus on battery electric for now since this segment was mentioned much more frequently on Twitter and it also has a slightly positive score. Donalson needs to catch the trend and expand their business in this field to maintain revenue and market share. Natural gas could be something Donalson can start to explore more and study on, relatively positive high sentiment score and lower frequency suggests natural gas is possible to be the next star in alternative powertrains.

Another suggestion is that Donadlson can reevaluate its supply chain and logistic channel. Our research on the article *The "Semiconductor Crisis"* indicates that pandemic and emergencies have high impacts on the powertrain market. Improvement on supply chain and logistic channels are able to help Donaldson increase efficiency of reacting on influences and reduce potential loss.