Donaldson
FILTRATION SOLUTIONS

Donaldson Project

NLP with Sentiment Analysis

About the Author

Team Bobcats

- **Mengyi Tao**
  tao.men@northeastern.edu
- **Zining Ta**
  tan.zi@northeastern.edu
- **Meiling Zeng**
  zeng.mei@northeastern.edu
- **Jiaqi Shen**
  shen.jiaqi1@northeastern.edu

Team Bobcats has four team members.
Project Manager: Mengyi Tao
Data Analyst: Mengyi, Zining, Meiling and Jiaqi

Table of Contents

1

Introduction

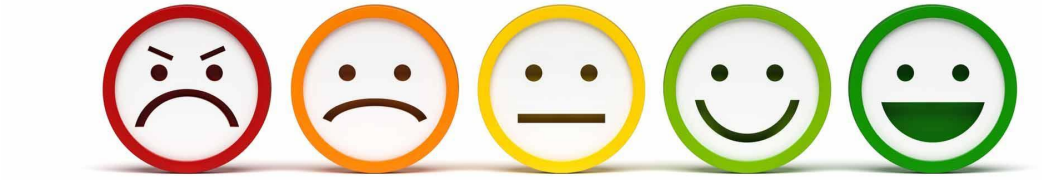
Introduction

Donaldson Company, Inc.is a global leader in the filtration industry, mainly engaged in the production and marketing of air filters (Donaldson, 2016). Its air filters are used in a variety of industries, including construction, transportation, aerospace, agriculture, food & beverage, chemicals, railroad, and pharmaceuticals.

Donaldson has been solving the complex filtering needs of its customers for more than 100 years. Today, they are one of the largest suppliers of unique filtration technology and high-quality filters and components (Donaldson, 2022).

Project Overview

For the organization's long-term growth, Donaldson wants to understand how the global market views different powertrains systems today. They hope to use a data perspective to understand which powertrains - battery electric, hydrogen fuel cell, H2 ICE, or any other technology - will gain the largest market share in the long term in order to adapt their product portfolio and strategy towards electrification. The project can also help them develop long-term strategies to maximize profits in the future. This is the main objectives of our project. Thus, we select YouTube platform to collect the relevant comments and conduct the sentiment analysis in order to help Donaldson understand the market sentiments towards different powertrains. Moreover, according to the sponsor's request, our project mainly focused on the truck and other heavy equipment market.



2

Analyzing Tools

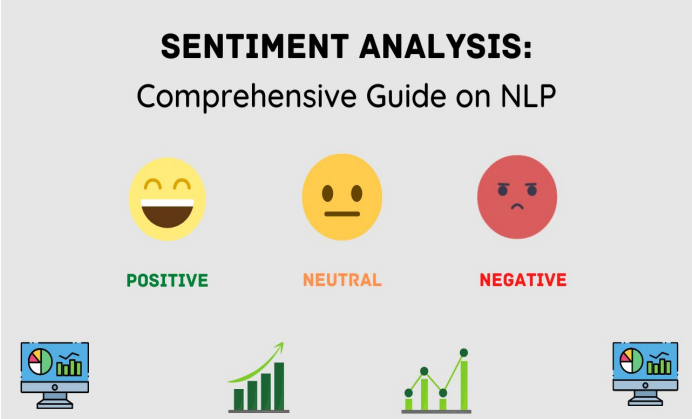
NLP & Sentiment Analysis

Intro for NLP

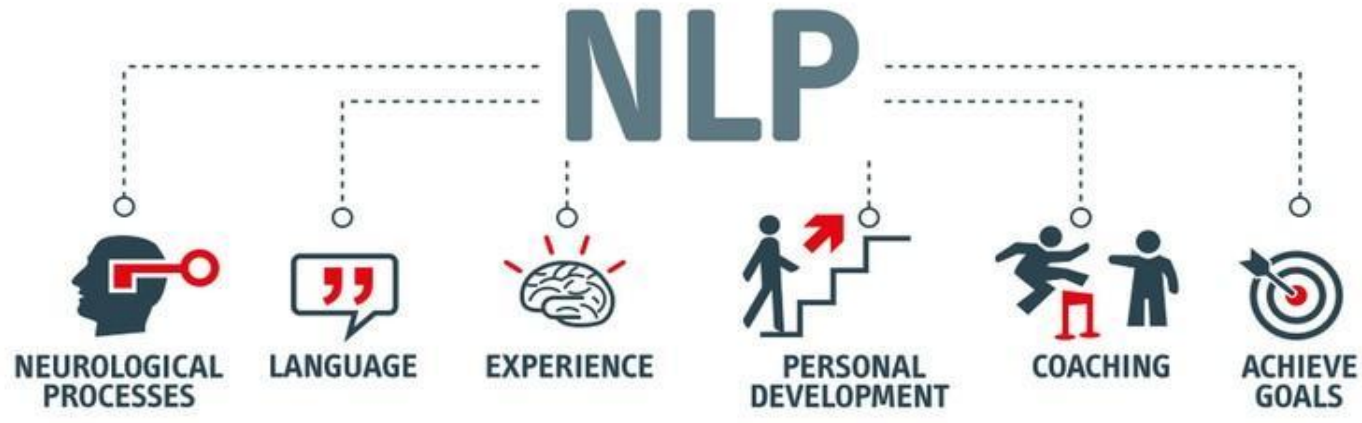
- Natural language processing (NLP) is the intersection of computer science, linguistics and machine learning.
- The field focuses on **communication between computers and humans in natural language** and NLP is all about making computers understand and generate human language.
- Applications of NLP techniques incorporate more than 10 fields such as language translation, text-filtering, speech recognition, sentiment analysis, voice assistants and chatbots, auto-correct and auto-prediction, etc.

What is Sentiment Analysis?

- Sentiment analysis is a natural language processing (NLP) technique used to **determine whether data is positive, negative or neutral from text data.**
- It can extract insightful information about the context.



Source: https://analyticslearn.com/sentiment-analysis-comprehensive-guide-on-nlp

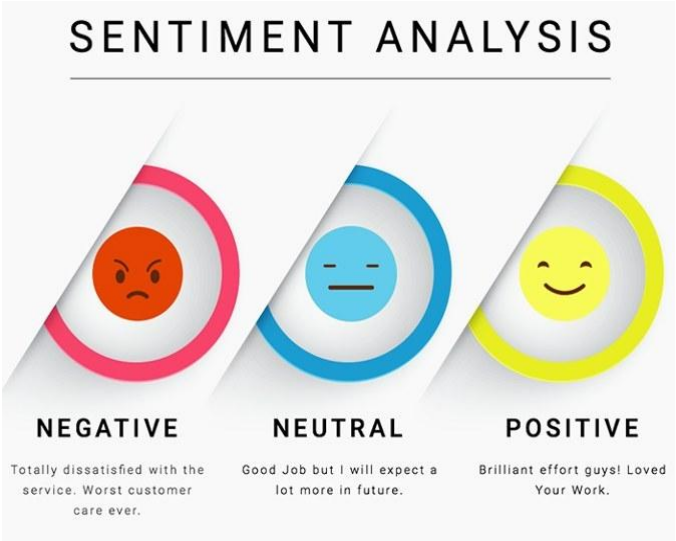


Source: https://www.shutterstock.com/zh/image-vector/banner-neurolinguistic-programming-nlp-vector-illustration-1257901324

Pros and Cons of Using Sentiment Analysis

Pros

- **Automatic and rule-based.** For unstructured data, it will automatically gather the text to analyze emotional tone as positive, negative, or neutral, and by their extent.
- **Quantified attitudes.** It can compute an algorithm that can give a score to each word and sentence. From the results, we can easily determine what the public is interested in and what they want to change.
- **Wide application for social media platforms.** Due to the rising popularity of social media, Sentiment mining from social media listening helps you analyze audience intent and opinions expressed on various social platforms.



Cons

- **Scene limitation of model itself.** How people express themselves in these domains is different, so when Sentiment analysis systems trained on review data, they're often much less accurate when applied to data from other domains such as news or social media
- **Sarcasm detection.** In sarcastic text, people express their negative sentiments using positive words.
- **Negation detection.** Negation is a way of reversing the polarity of words, phrases, and even sentences. It's difficult to use different linguistic rules to identify whether negation is occurring
- **Word ambiguity.** The impossibility to define polarity in advance because the polarity for some words is strongly dependent on the sentence context.
- **Multilanguage limitation.** The accuracy of analyzing non-English language will drop due to inaccurate translation.

Source: https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/

3

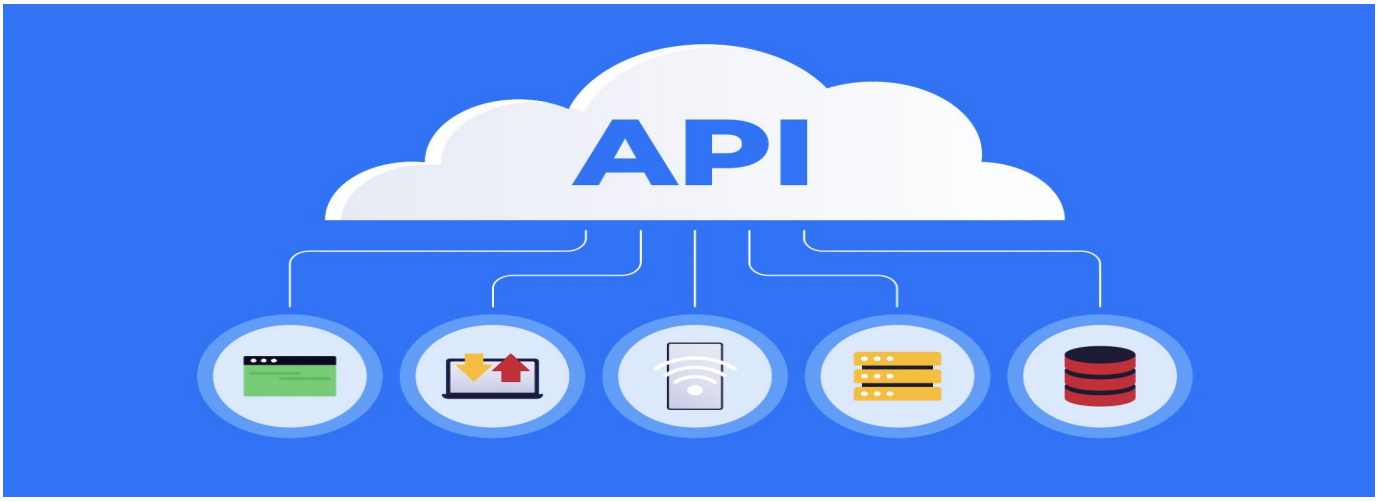
Data Collecting

API Intro

Application Programming Interface (API) is a way for two or more computer programs to communicate with each other. It is a type of software interface, offering a service to other pieces of software (Reddy, 2011). A document or standard that describes how to build or use such a connection or interface is called an API specification. Computer systems that meet this standard are referred to as implementing or exposing APIs. The term API can refer to a specification or an implementation.

APIs are one of the more common ways technology companies integrate. The main policies (Mark, 2014) for releasing an API are:

- **Private**: The API is for internal company use only.
- **Partner**: Only specific business partners can use the API.
- **Public**: The API is available for use by the public.



Source: https://appmaster.io/blog/apis-for-beginners-how-to-use-an-api-a-complete-guide

YouTube Video Comments

YouTube Platform Benefits:

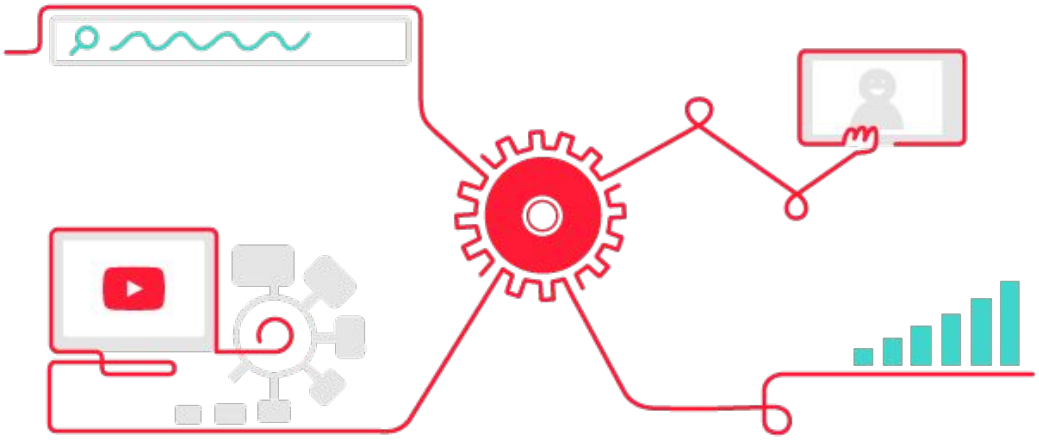
- **Official YouTube Account:** Many targets heavy equipment company have an official YouTube account so we can search related videos about new power engine or powertrain of truck though those accounts.
- **Keywords:** We select several keywords which interest Donaldson, like: Battery, battery-electric truck, hydrogen, hydrogen truck, H2 ICE, Fuel cell truck and other new clean power.

YouTube Video Selecting rules:

Restricting rules of selection will improve the quality of the videos and sentiment analysis of comments.

- **Watches: over 10k**
- **Time: In the last 2 years**
- **Number of comments: over 50**

Source:
https://www.youtube.com
https://www.youtube.com/watch?v=45TicSq1NGg



YouTube Data API

YouTube Data APIs:

With the YouTube Data API, we can easily use api-key to get access all the comments of target YouTube videos and transfer those comments as a raw data set.

Benefits: easy to assign and get access

First, we use our own api-key:

```
api_key = "AIzaSyADsZcuJRGJ0kMS_XOJ-HqVcNvtKn6XdsI"
```

Second, we use the video-id:

```
ID = "tAhnzugkZv0"
```

Third, we define a function called:

```
def scrape_comments_with_replies(ID):
```

This function can use the video-id and api-key to get access the comments of the YouTube video.

Source:https://developers.google.com/youtube/v3

Forth, then we turned those comments into a csv file.

```
df.to_csv('youtube-comments.csv'
```

The comment line in csv file looks like this:

Comment
Aa Chal ke tunea8 didnemeu ejebe
Only 25999
Im getting a new 85kw put in my 201
The budget is 5 per video a Very Chea

We have nearly 110k lines of comments from 135+ videos.

Finally, when we downloaded the comments, there were some punctuation, emojis, stopwords and messy word in those data.We did data cleaning in the next part which would help us understand more about the data and improve the efficiency and accuracy of the sentiment analysis.

4

Data Cleaning

Data cleaning

Data Cleaning Steps:

Decoding and Encoding:

- Interpret comments as ASCII code.
- Code:

```
df['Comment'] = df['Comment'].str.encode('ascii','ignore')
df['Comment'] = df['Comment'].str.decode('ascii','ignore')
```

- Tips: ASCII (Mackenzie,19080), stands for American Standard Code for Information Interchange. It's a 7-bit character code where every single bit represents a unique character.

Source: https://www.inzata.com/what-is-data-cleaning/

Remove punctuation:

- The punctuation removal process will help to treat each text equally. We need to take care of the text while removing the punctuation because the contraction words will not have any meaning after the punctuation removal process. We also need to be extra careful while choosing the list of punctuations that we want to exclude from the data depending upon the use cases.
- Code:

```
df['Comment'] = df['Comment'].str.replace('[^\w\s]','')
df['Comment'] = df['Comment'].str.replace('[\n]','')

df['Comment'] = df['Comment'].str.replace('[_]','')
```

- Input:

```
'!hi. wh?at is the weat[h]er lik?e.'
```

- Output:

```
hi what is the weather like
```

Source: https://www.educative.io/blog/what-is-data-cleaning

Remove Stopwords:

- Stop words are available in abundance in any human language. By removing these words, we will give more focus to the important information.
- Code:

```
stop = stopwords.words('english')
df['new_reviews'] = df['Comment'].apply(lambda x: " ".join(x for x in x.split() if x not in stop))
```

- Input:

```
'Nick likes to play football, however he is not too fond of tennis.'
```

- Output:

```
Nick likes play football , fond tennis .
```

Source: https://www.grantbook.org/blog/how-to-clean-your-data-hygiene-best-practices

Replace Emojis into Words

- Nowadays in a day to day life, people often use emoji and emoticon in a sentence to express the feeling or describe object instead of writing a word in a social media platform. Sometimes, emoticons from emojis give strong information about a text such as feeling expression. So we replace emojis into words during data cleaning.

- Code:

```
with open('/content/Emoji_Dict.p', 'rb') as fp:
    Emoji_Dict = pickle.load(fp)
Emoji_Dict = {v: k for k, v in Emoji_Dict.items()}

def convert_emojis_to_word(text):
    for emot in Emoji_Dict:
        text = re.sub(r'('+emot+')', "_".join(Emoji_Dict[emot].replace(",","").replace(":","").split()), text)
    return text
```

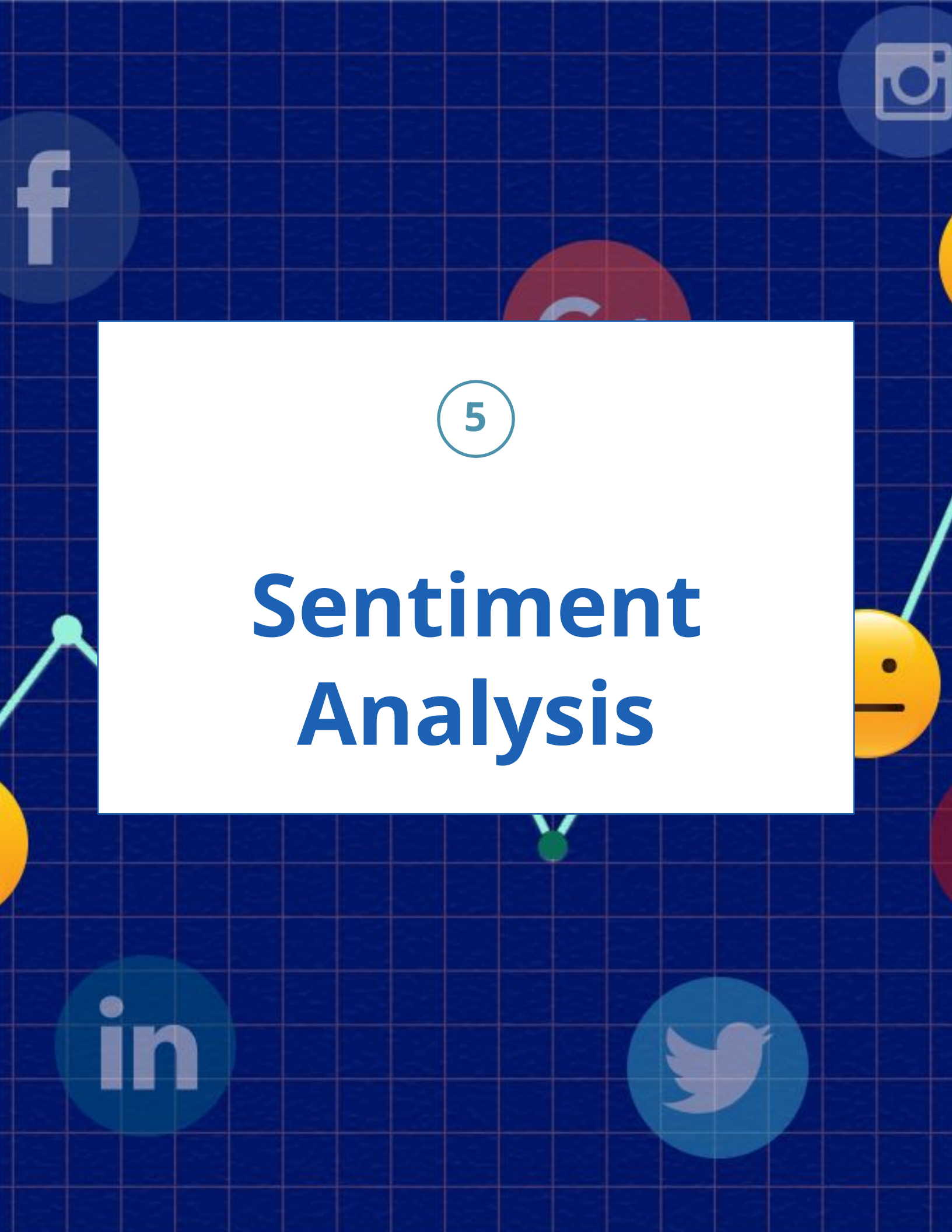
- Input:

```
I am 😊
```

- Output:

```
I am smiling_face_with_smiling_eyes
```

Source: https://www.educative.io/blog/what-is-data-cleaning


5
Sentiment Analysis

Sentiment Analysis

Sentiment Analysis Steps:

We use the Affin and NLTK package to score each comment by adding the sentiment scores of all the words in the comment. If the final score is positive, then this is a positive comment, or the comment would be marked as negative comment. We will compare the results with the real sentiment and build a model to predict the real sentiment with the results we gain from Affin and NLTK package.

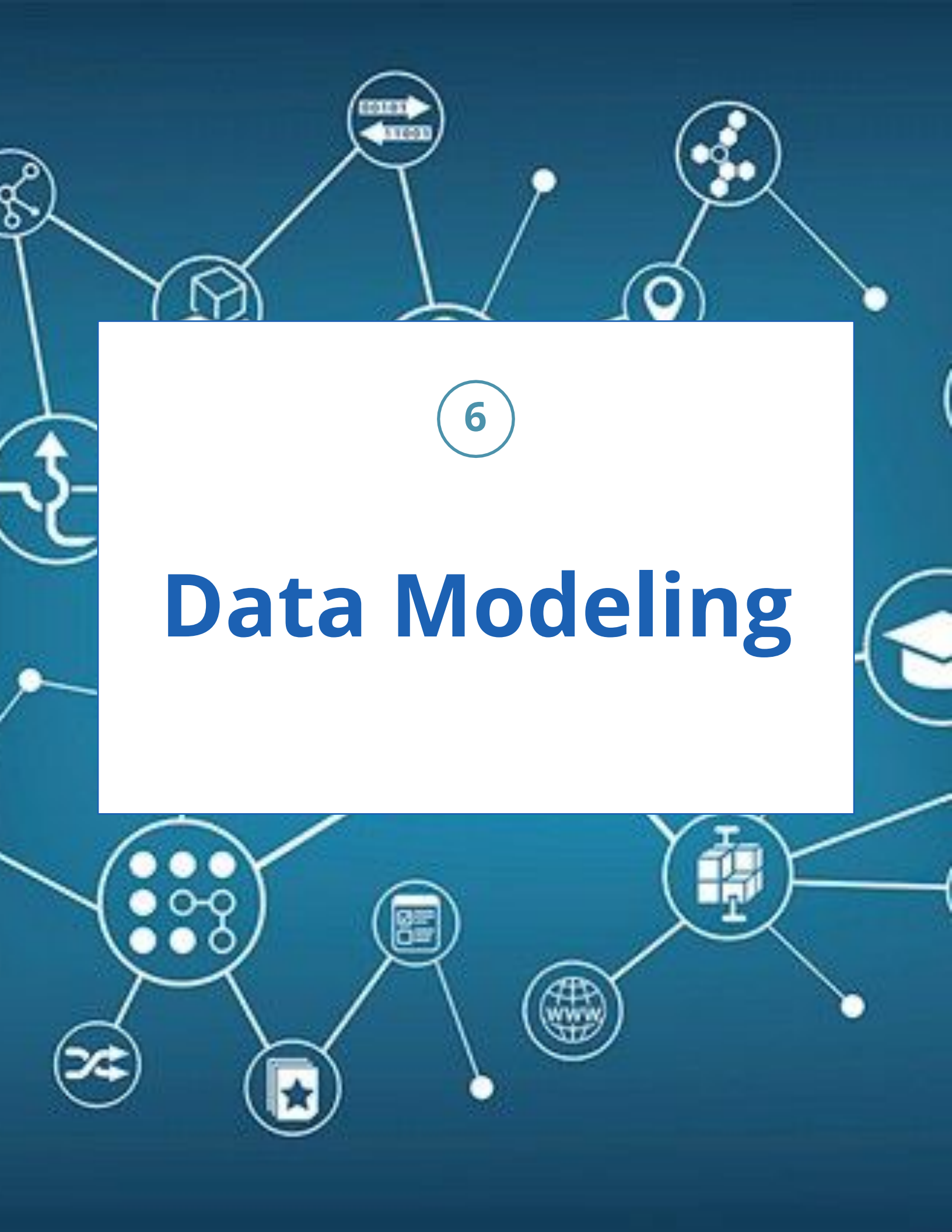
Affin Package:

Affin is one of the most popular lexions used to perform sentiment analysis developed by Finn Årup Nielsen. It contains over 3300 words and have a polarity score associated to each word. The score goes from -5 to 5.

NLTK Package:

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet. The online version of the book has been updated for Python 3 and NLTK 3 (nltk.org, 2022).

keyword	Comment	affin_sentiment	nltk_sentiment	affin_score	nltk_score	Real_Sentiment
electric	updatewe love your show spacex We travelec	pos	pos	6	0.8555	pos
electric	Tesla charge 20k to replace So technology isnt r	neu	neg	0	-0.3343	pos
electric	Lithium price 216 500CNYt	neu	neu	0	0	neu
electric	Well the Electric is the engine so to speak its als	neg	neg	-8	-0.5131	neg
electric	The slogan Truth to Power is really gay	neu	pos	0	0.3182	pos
electric	not to understand fully	neu	neu	0	0	neu
electric	Well you have got to leave some things for the t	neg	pos	-1	0.2263	neu
electric	Oky sooo whenever musk says bullshit like 10x le	neg	neg	-3	-0.0092	neg
electric	Good job Tesla	pos	pos	3	0.4404	pos
electric	The axiomatic ton naturally bless because salesr	pos	pos	3	0.8074	pos



6

Data Modeling

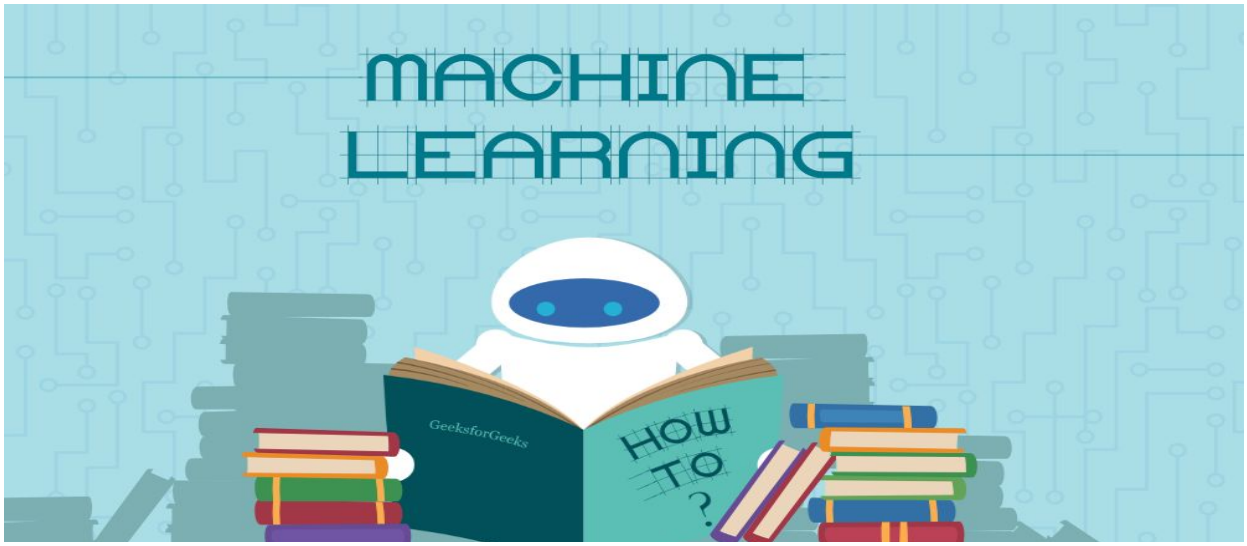
Data Modeling

Logistics:

We use keywords, score, and result of the two nlp packages to build a machine learning model to predict the real sentiment of a comment. For doing classification in Python, we labeled the real sentiment column as 1, 2 ,and 3 and we created dummy variables for categorical variables in our dataset.

Classifiers:

- **Random Forest:** An ensemble learning method for classification tasks that operates by constructing a multitude of decision trees at training time.
- **Decision Tree:** A decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
- **Gradient Boosting:** Gradient boosting (Piryonesi, 2020) is a machine learning technique used in classification tasks which gives a prediction model.



Source: https://www.geeksforgeeks.org/machine-learning/

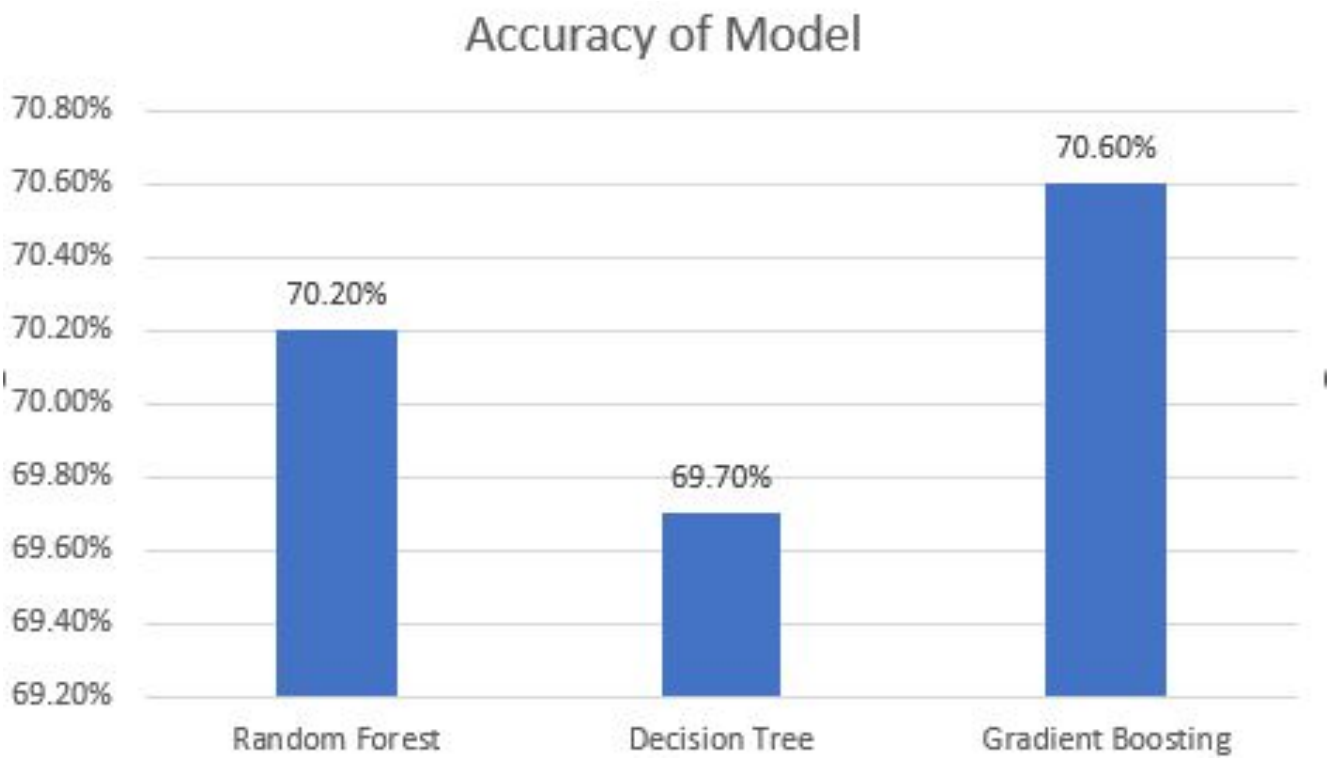
Data Modeling

Estimate indicators:

We use the accuracy to estimate our model. The accuracy is calculated by the percentage that the predicted results is the same as the real value in the test dataset.

Result:

The accuracy for random forest , decision tree, and gradient boosting is 70.2%, 69.7%, 70.6%. We can conclude that gradient boosting have the best performance on our dataset.



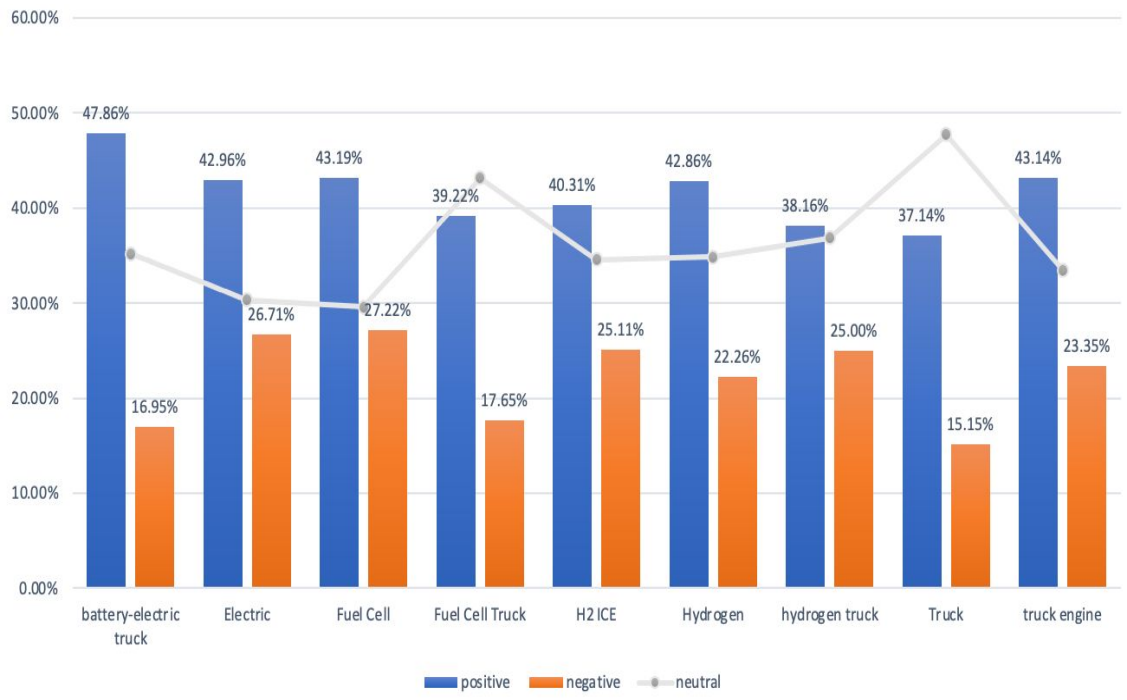
7

Data Visualization

Data Visualizations

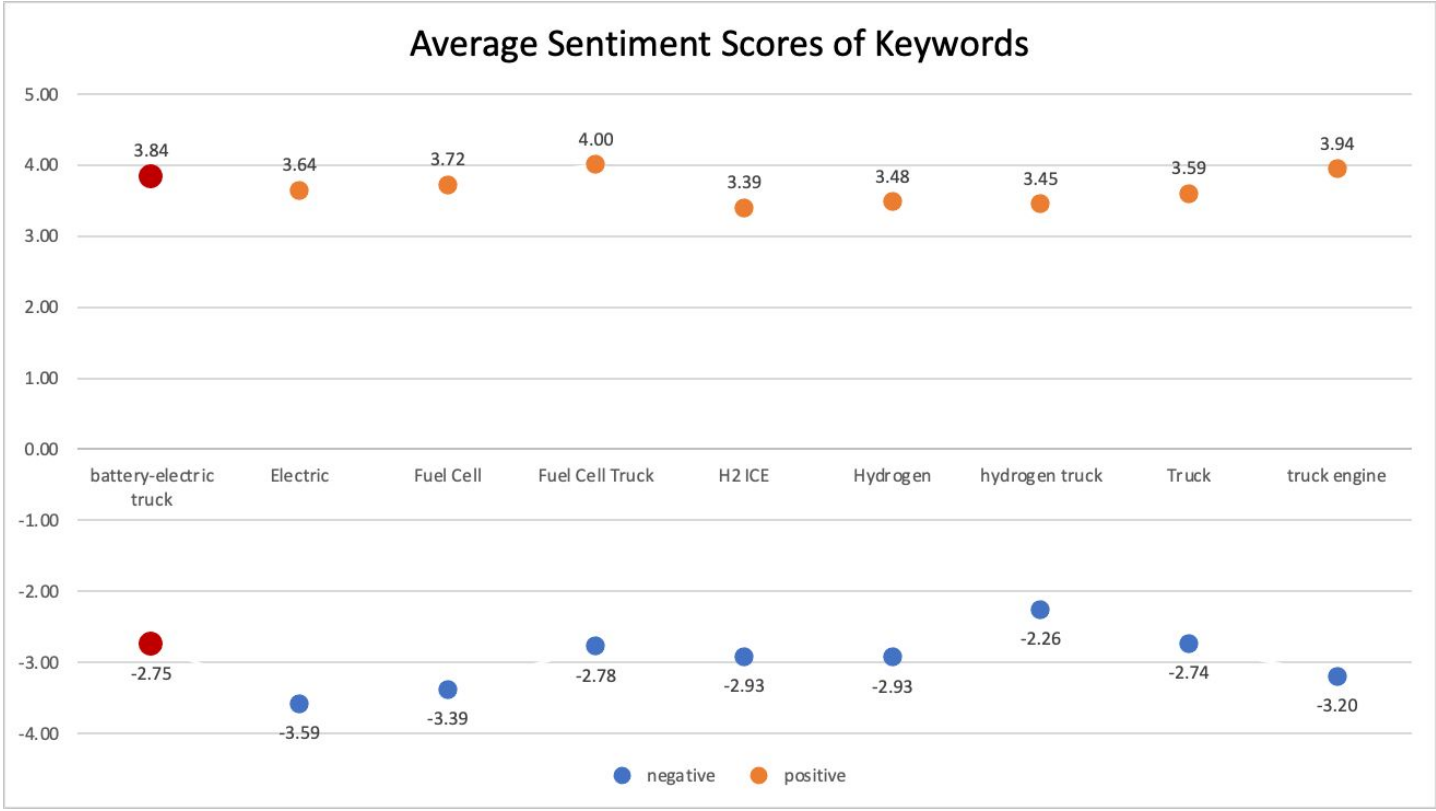
How to analyse those plots and what will be shown?

Compare plots between positive, negative, neutral attitude of keywords will show that which keyword has the most positive percentage and less negative percentage, which keyword is the most controversy, the relationship between the new clean power, powertrain with truck engine. These will help us understand not only the powertrain market but also the powertrain of truck market.



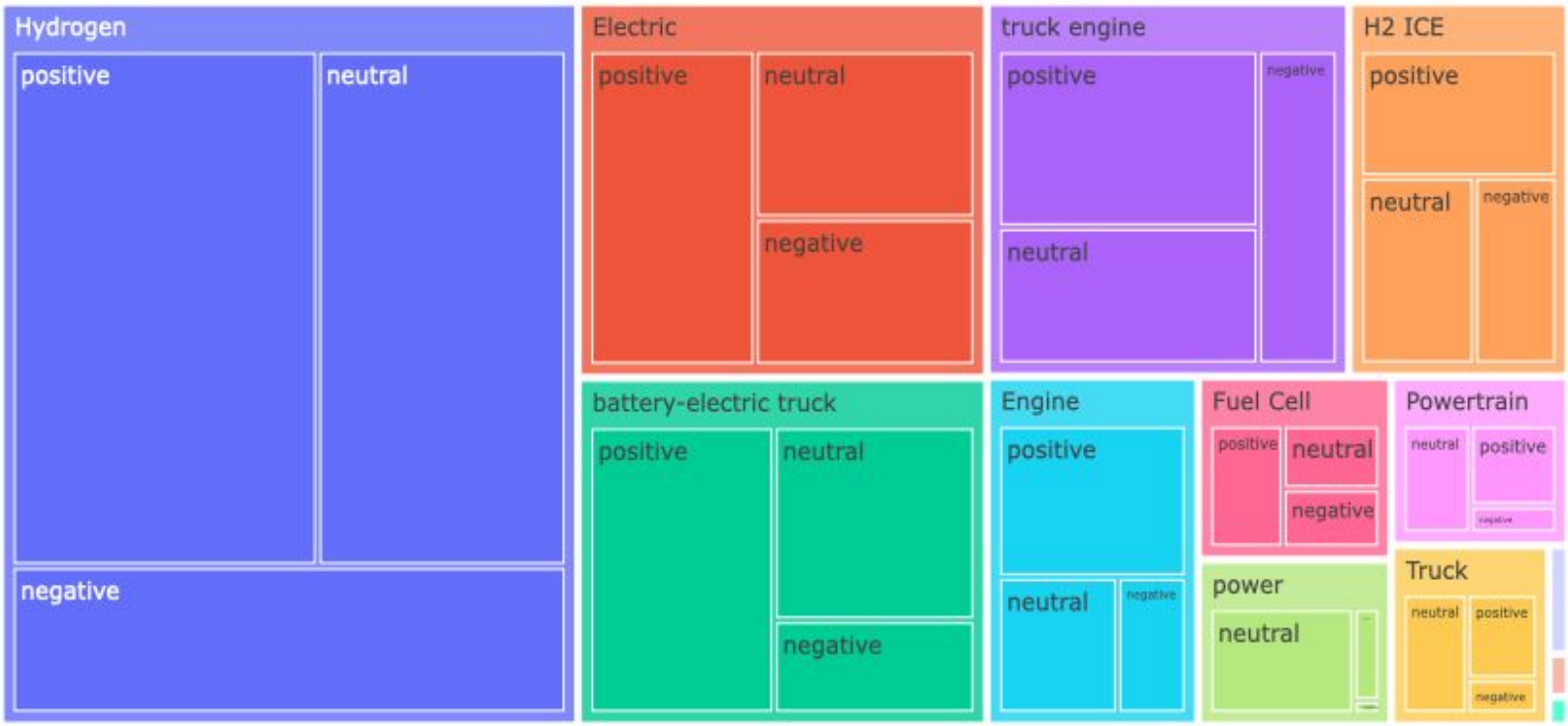
We can see from the barplot that most keywords have more positive comments than negative comments. Therefore, we can find that most people who are interested in these keywords respond positively and battery-electric trucks may be a winning trend in the future.

Data Visualizations



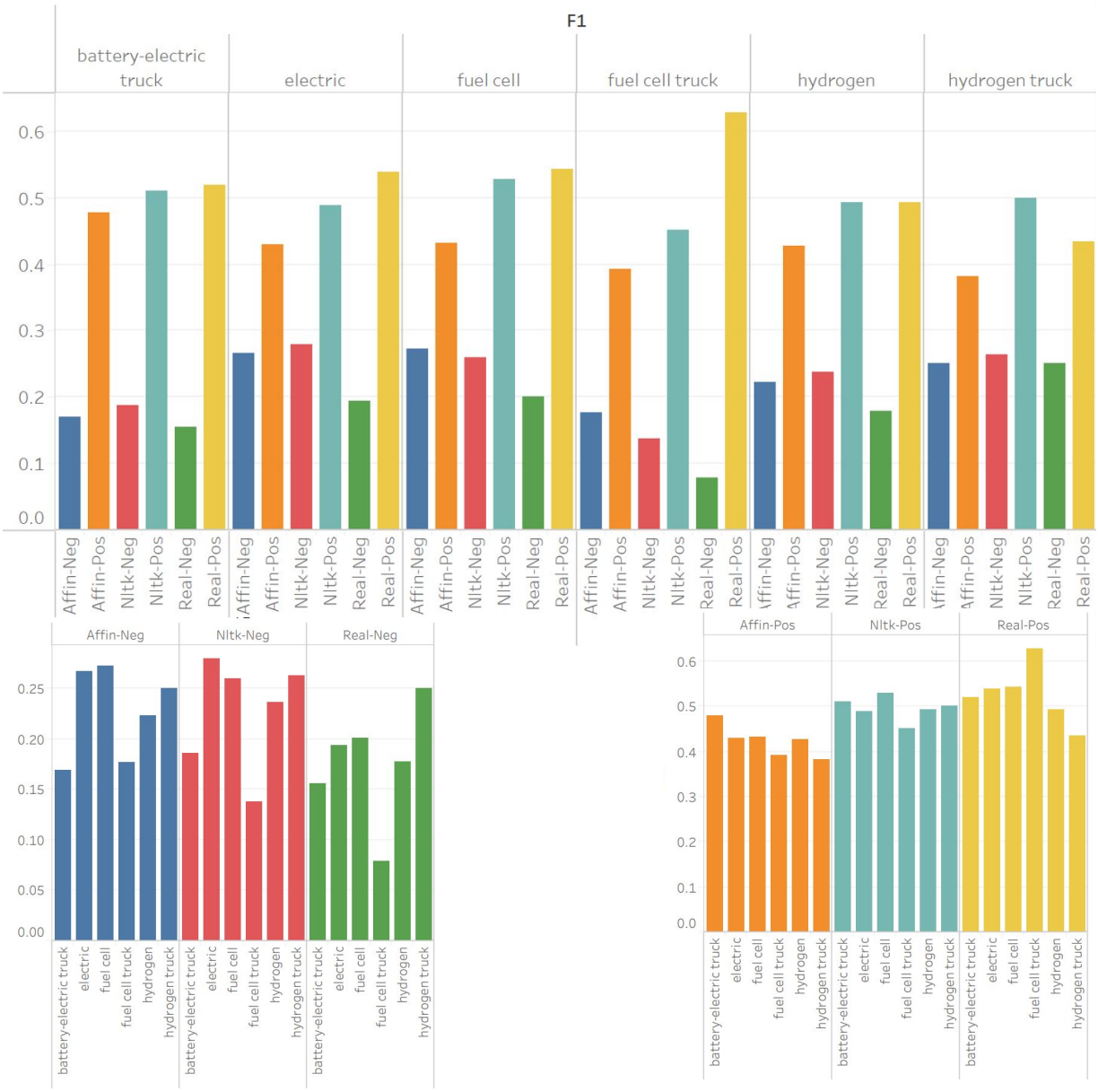
From the scatter chart, we can see that both positive and negative comments for some keywords like “battery-electric” have high average scores, which means people's attitudes Donaldson understands the trend of new power powertrain and engine of the truck industry with the winning technology towards these keywords are controversial.

Data Visualizations



According to this chart, we can clearly find that hydrogen occupies the biggest area in the map which means most people are interested in this type of fuel, however the hydrogen truck does not attract people's attention. Otherwise, they pay attention to the electric and battery-electric trucks in the same level. This kind of situation will help us and Donaldson understand the popular fuel in the truck powertrain and engine industry with the winning technology in the future.

Data Visualizations



According to the bar chart, we can conclude that between the two packages we used for sentiment analysis, the result of the NLTK package is more close to the real sentiment. Among all the keywords, fuel cell truck have a higher percentage of positive sentiment in the real sentiment contrary to the Affinn sentiments and NLTK sentiments.


8
Conclusion

Conclusion & Suggestion

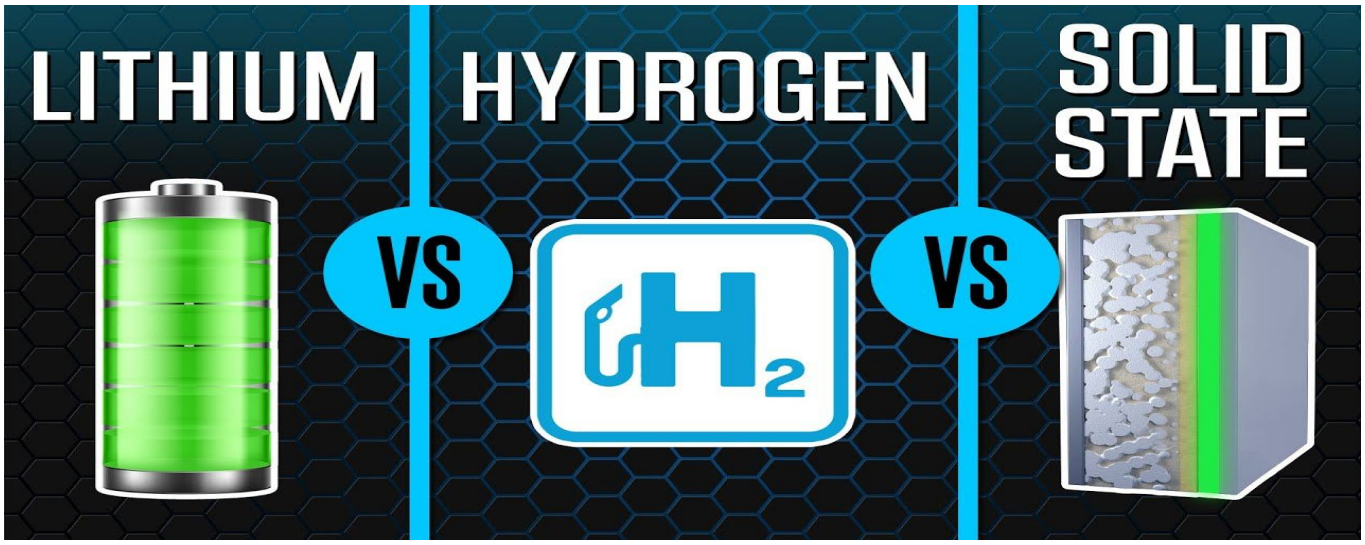
Conclusion:

We prefer to select the battery-electric powertrain as the winning technology of powertrain industry in the future. Based on the sentiment analysis outcome, we find that people are more interested in the battery-electric truck and hydrogen truck than other new clean power. They demonstrated a positive attitude towards the future of the battery-electric powertrain.

However, when they talk about hydrogen powertrains and trucks, they have a controversial discussion about hydrogen as an energy source because the current market for such powertrain systems is not in large numbers. Therefore, people hold contrary opinions towards hydrogen engines.

Suggestion:

- Donaldson could prioritize the development of battery-electric truck powertrain filters, with the goal of occupying the mainstream market for truck powertrain filters in the future.
- Distribute some production lines and machines to produce filters that can filter hydrogen, H2 ICE and other fuel cell energy sources, with the aim of keeping abreast of market changes and effectively making strategic changes.



Source: https://www.youtube.com/watch?v=CcRAEAtRlFE

Weakness & Suggestion

Sentiment analysis is just one way to do the NLP, and it has some limitations, such as word ambiguity, sarcasm detection and scene limitation of model itself. Besides, there are also some weaknesses in our project.

- The number of effective videos under each keyword might cause bias and affect final results. If the number of videos of a certain keyword outnumber the others, the result might show that people have greater interests in this field over others.

- If our team only used a single python sentiment analysis package for this project, it may lead to unconvincing analysis results. However, if we use multiple packages to compare the results, since the criteria and dictionary of each package are different, it will cause the final results to be biased. Therefore, using data modeling to find the relationship between package outcomes, keywords and real sentiments is of little practical significance.

- Collecting and analyzing the comments on social platforms to draw conclusions is not enough, especially YouTube. Because many heavy equipment companies don't have official accounts on YouTube or rarely post YouTube feeds and videos. Moreover, lots of comments on social platforms can only represent for the attitude of some people who are not the companies' target customers. They are not going to buy trucks or other heavy equipment. Therefore, our team and Donaldson should consider looking at other platforms and using sentiment analysis to analyze, such as reports or papers published by professional consulting firms.

Reference

- Donaldson (2022). Advancing filtration for a cleaner world Available at: https://www.donaldson.com/en-us/about-us/who-we-are/at-a-glance/
- Reddy, Marathi (2011). *API Design for C++*. Elsevier Science. p. 1. ISBN 9780123850041.
- Boyd, Mark (2014-02-21). "Private, Partner or Public: Which API Strategy Is Best for Business?". *ProgrammableWeb*. Retrieved 2 August2016.
- NLTK Org (Mar 25, 2022) https://www.nltk.org
- Mackenzie, Charles E. (1980). *Coded Character Sets, History and Development* (PDF). *The Systems Programming Series* (1 ed.). Addison-Wesley Publishing Company, Inc. pp. 6, 66, 211, 215, 217, 220, 223, 228, 236–238, 243–245, 247–253, 423, 425–428, 435–439. ISBN 978-0-201-14460-4. LCCN 77-90165. Archived (PDF) from the original on May 26, 2016. Retrieved August 25, 2019.


Thank you!
Fuel Filter