Electric vehicles (EVs) are powered by electricity from batteries or fuel cells, rather than gasoline or diesel. The concept of an electric vehicle was introduced in the early 1990s. However, the technology was not practical at that time. Nowadays, the technology among batteries, motors, and chipsets become more advanced enough to produce the vehicle in the range of 300-350 miles per one charge, which allows many people to commute between their home and their nearby destinations. Today, there are several types of electric vehicles: all-electric vehicles (EVs), hybrid electric vehicles (HEVs), and plug-in hybrid electric vehicles (PHEVs). These vehicles have become increasingly popular in recent years due to their potential to reduce dependence on fossil fuels and decrease emissions .

A collage of a car

Description automatically generated with low confidence

The EV type

“Do you prefer to use EVs over gas-powered vehicles?” is the main idea of this text analysis project. After gathering text data from various sources, there are evidently two sides: people who support buying and using the vehicle and people who tend to stick with gas-powered vehicles. Regarding the supported group, they love the way the EVs produce zero emissions while in operation. This helps to reduce air pollution and improve public health. Air pollution has been linked to a wide range of health problems, including respiratory issues, heart disease, and stroke. By reducing the number of emissions-producing vehicles on the road, EVs help to improve air quality and reduce the health impacts of air pollution. In addition, the EVs fans think that EVs can also help to reduce dependence on fossil fuels and decrease the greenhouse gas. This means the vehicle emits less greenhouse gasses such as carbon dioxide and carbon monoxide, which help relieve the global warming issue.

For the EVs opposers, they believe that the production of EVs and their batteries require the mining of rare earth metals such as lithium, which has a more detrimental effect on the environment. Some components in the battery cannot be recycled and are a risk to people's health. Another concern of opposers is the lack of infrastructure. Many people are likely to use an EV as their primary cars in daily life and road trips. The problem of the lack of charging points causes them not to be able to use them on their road trip. Some consumers do not have their own charging area and charging facility. This makes it difficult to use an EV on a daily basis. In addition, the range of battery is not enough if the EVs are used at the freeway’s speed. The batteries tend to reduce very fast. Finally, the cost of EVs is still relatively high compared to traditional gasoline-powered vehicles, which can make them less accessible to some consumers.

A picture containing several

Description automatically generated

The traffic in a big city

In the US, many reports indicate that 91.55% of households will have access to at least one vehicle in 2020, which is up from 90.82% in 2015. The estimated number of cars is around 270 million in 2020, which is up from 260 million in 2015. This is just for a country. If we imagine how many cars there are in the world, the number could be more than 20 times the number of cars in the US. The more cars we have, the more air pollution will exist. Household combustion vehicles are common sources of air pollution. Pollutants of major public health concern include particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulfur dioxide. Outdoor and indoor air pollution cause respiratory and other diseases and are important sources of morbidity and mortality. In addition, the combustion of the vehicle can cause fine particulate matter (PM2.5) in the air. Fine particulate matter (PM2.5) is an air pollutant that is a concern for people's health when levels in air are high. PM2.5 are tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. Previous epidemiological studies have indicated that ambient PM2.5 is carcinogenic, and may increase the morbidity and mortality rates associated with lung cancer. Therefore, renewable energy cars such as EVs are the potential solution to help solve this situation in the long term plan because they can help limit the air pollution into the environment.

A person standing in the middle of a busy street

Description automatically generated with low confidence

PM2.5

In addition to concerns about air pollution, this report aims to deliver more about the problems of EVs that have impacts on both the environment and consumers. There are problems with the infrastructure network: EVs’ cost and the battery components as mentioned previously. Sustainable development is the key for product development. The idea of sustainable development is to provide necessary resources and ecosystem services to humans without the environmental damages. The major issues can be solved if all EV’s manufacturers are concerned about the issues and feel determined to solve them. Besides, all countries have to find enough energy sources for their vehicles. Currently, most countries use fossil fuels to produce electricity, which does not go along with the purpose of the EVs. More research in the material sciences to improve solar cells and battery efficiency should be considered a priority. The sunlight is a free and clean energy source. If there is a way to produce more electricity from sunlight and its eco-friendly storage, many opponents may turn into supporters in the future.

A factory with smoke coming out of it

Description automatically generated with low confidence

The fossil fuel emission

Both news and users’ opinions are collected from Reddit, Twitter, NewsAPI, CNBC and InsideEVs. The first two sources are about users’ opinion and discussion. NewsAPI, CNBC and InsideEvs are the news and technology in EVs. Reddit, Twitter, NewsAPI data are collected via the APIs, whereas the two latter sources are from web scraping. After all data has already been collected, they are classified into five labels according to the problems of EVs in the previous paragraphs. The labels are environment, news, battery, infrastructure, and cost. The label of environment talks about the impact of EVs on the environment; the label of news explains the information and technology in EVs; the label of battery is the battery capacity, charging time, and the battery’s efficiency; the label of infrastructure is the charging networks and its facilities; the last one is the cost of the EVs and electricity fuels in different locations. All special characters are removed from the text data and transformed to lemmas by lemmatization and stemming techniques. After that , the data is used to create a bag of words by the CounterVectorizer package, which helps to remove the sequences between terms in documents. However, CounterVectorizer cannot be used to measure the importance of terms among all documents. So, Term Frequency-Inverse Document Frequency (TF-IDF), is employed to measure how important a term is within a document relative to a collection of documents.

After the text data is already prepared in the desired format, clustering is also implemented in this project. Clustering is an unsupervised learning technique used in machine learning to group together similar data points based on their features. The goal of clustering is to find patterns in the text data that can help to better understand the underlying structure and relationships between the texts. Several techniques such as K-Means clustering, Hierarchical clustering, Association Rules Mining, and Latent Dirichlet Allocation (LDA) have been used in this project. K-Means, Hierarchical clustering, and Association rules mining indicate separating into five clusters is the suitable clusters based on the results. The five main discussions in this dataset are EVs’ news, EV’s price, charging infrastructure, environmental impacts from battery production, and Elon Musk’s roles in the EVs’ society. However, LDA shows that only three clusters are enough to explain the discussions within this text data. The three main topics are battery properties and attributes, Elon Musk and Tesla, and environmental impacts from EVs’ batteries.

A picture containing text, clipart, screenshot

Description automatically generated

Unsupervised Learning

Every method uses different metrics to help select the proper number of clusters. Silhouette score is used to help select the number of clusters in K-Means. The score will be compared between the distance between the same predicted cluster and the nearby predicted cluster in every data point. Dendrogram will help visualize how close between subtopics are and the results can help us to select the proper clusters in the Hierarchical clustering. Several graphs are used to visualize rules from the Association Rules Mining and the number of clusters can be roughly estimated from the graphs. In LDA, coherent metric and pyLDAvis are used to help score and visualize the results. The outputs are the related words in the abstract topic. After the words are scored by the coherent metric, the higher score indicates that clusters are easier to understand, which helps to choose the proper number of clusters.

As mentioned before, some of the data have been labeled in the five categories. The data can be used to perform the text classification. The supervised learning methods will be implemented because the methods rely on the prior knowledge of what the output values for our samples should be. Hence, the goal of supervised learning is to learn a function that, given a sample of data and desired outputs, best approximates the relationship between input and output observable in the data. In this project, several supervised learning methods such as Naïve Bayes, Decision tree, and Support vector machines (SVM) are implemented to assess the accuracy on the testing dataset. Among all models, the SVM’s accuracy outperforms the accuracy from the other models. The SVM’s accuracy is around 77.0%, whereas the decision trees and Naïve Bayes accuracy are around 65.0% and 74.0%. However, the most suitable method may be Naïve Bayes because the model uses less time-processing with high accuracy. Before being fed into the models, all label terms are removed from the documents to avoid bias during the training process.