Final Report

Cardiff University School of Computer Science and Informatics

ANALYSING THE DISCOURSE RELATED TO ELECTRICAL VEHICLES ON SOCIAL MEDIA

CM3203 – One Semester Individual Project – 40 Credits

Author: Anthos Kountouris

Supervisor: Dr Padraig Corcoran

Abstract

The electric vehicle (EV) revolution is speeding up as, according to a research of Canalys (2022), a global technology market analyst firm, global electric car sales grew 109% and 6.5 million EVs were sold worldwide in 2021. Also, by 2035, the UK government has committed to banning the sale of new vehicles that are solely powered by petrol and diesel and transitioning to electrical vehicles (Pickett et al. 2021). Many people support this goal and believe that it will result in a more sustainable world. On the other hand, some people have concerns about this transition. These concerns include not being able to buy new electrical vehicles, but also not being able to charge them.

The growth of the internet coincides with the increased use of social media platforms, such as Instagram, Twitter and Facebook, which improves the efficiency and convenience of communication around the world. Twitter, which is considered one of the leading social networks, allows people to communicate and express their thoughts in short, 280-character posts, also known as ‘tweets’.

In this project, Natural Language Processing techniques such as Sentiment Analysis and LDA Topic Modeling had been used, in order to analyse the sentiment that people have and the topics that are being discussed on Twitter regarding electric vehicles. The results showed that the positive sentiment of people was higher than the negative. Some of the major findings indicate that sustainability may be one of the main reasons why people support the transition to EVs, and that mining for EV battery materials and charging stations challenges might be some of the main factors that hold people back from buying an EV. Additionally, it has been observed that real-time events might have also affected people’s sentiment towards EVs.

Acknowledgements

I would like to thank my supervisor Dr Padraig Corcoran for his guidance and feedback throughout this project. I would also like to thank my family, especially my sister, and friends for their support during the compilation of this project.

Table of Contents

Table of Figures

List of Tables

1. Introduction

The transition from fossil-fuel vehicles to electric vehicles is coming sooner than we thought. According to the UK government, by 2030 the sale of new vehicles that are solely powered by petrol and diesel will be banned in the UK, in order to encourage the uptake of zero-emission cars and vans (Pickett, L. et al. 2021). In 2021, the US automobile giant General Motors announced that they plan to stop selling petrol-powered and diesel models by 2035. It is a fact that the governments of many countries will accelerate change. According to BloombergNEF (BNEF) consultancy in London, in 2035 half of global passenger-vehicles sales will be electric even with no new regulations or policies (Castelvecchi 2021).

In order for this transition to be achieved successfully, the number of automotive batteries produced every year will need to be increased. Many European countries are investing to create economic growth around battery manufacturing for EVs and they are making plans for new manufacturing centres in continental Europe to see 450 GWh annual battery production capacity by 2030 (Pickett, L. et al. 2021). But while electric vehicles can play a critical role in the reduction of emissions in the transportation sector, they still come with environmental costs. A huge concern is the supply of Lithium, Cobalt, and other materials that the batteries need in order to be built. For example, some of the proposed sites for mining Lithium are located in historic Indigenous lands. Also, Lithium mining can be water, land and energy-intensive. Mining these materials not only affects the environment, but the humans as well. There have been some reports about low safety conditions in mines and some others about human rights violations, as children are being forced to mine (Climate Nexus [no date]).

Another concern about electric vehicles is the charging stations. There is currently some uncertainty as to how many charging stations are needed and where they should be located. EV owners rely mostly on workplace or home charging, but they want to have more extensive and fast public charging, to have the opportunity to make longer journeys with their vehicles. The fear of not knowing if the car will make it to the next charging station, also known as “range anxiety”, is one of the reasons that people are holding back from buying electric vehicles (Pickett, L. et al. 2021).

There are many concerns and different opinions about electric vehicles. It is a topic that has attracted a lot of people the last few years. Some people support the goal of the transition and believe that it will result in a more sustainable world and other people have their concerns.

This project focuses on analysing the discourse related to electrical vehicles on social media and specifically Twitter. Social media is a very well-known concept to people nowadays, it has grown rapidly, especially in the last decade and according to Datareportal, 4.62 billion people around the world now use social media (Kemp 2022). People are communicating through social media but also expressing their concerns and thoughts. Twitter, which is one of the most popular social media platforms, allows people to interact and share their opinions and thoughts about anything, through the format of short messages of 280 characters, referred to as ‘tweets’. Many research studies are using Twitter as a data source, as it provides its data via a number of Application Programming Interfaces (API).

The research goals of this project are to find the sentiment that people have and the topics that are being discussed regarding electric vehicles on Twitter, but also to analyse and interpret these topics to understand whether they affect people’s sentiment or not. In order to achieve

these, Natural Language Processing techniques will be used, such as Sentiment Analysis and Topic Modeling. It is important and valuable to understand how people feel about electric vehicles, because the process of transition is already ongoing.

2. Background

In this section, I will introduce and discuss the most important concepts and tools that I will use in this project, along with their functionality.

2.1 Twitter API

The Twitter API allows you to read and write twitter data. It can be used to compose tweets, read profiles, and access followers' data and a high amount of tweets on particular topics in specific locations. In the case of this project, it will be used to retrieve tweets that are related to electric vehicles. In order to access the Twitter API, a Twitter 'Developer Account' is needed. Once an account is created, unique authentication credentials are generated which can be used to retrieve data from Twitter without the need of accessing the website manually. The 5 unique credentials generated are: 'Bearer Token', 'Consumer Key', 'Consumer Secret', 'Access Token' and 'Access Token Secret'.

2.2 Tweepy

As the programming language I will be using in this project is Python, I chose the library called 'Tweepy', which is an easy-to-use Python library for accessing the Twitter API. It can be used to establish a connection with Twitter using Twitter's API unique credentials, and retrieve data. The Twitter tweets gathered using Tweepy are rendered in JavaScript Object Notation (JSON) (Rude 2022). JSON is a file format that uses human-readable text to store and transmit data objects consisting of attribute-value pairs and arrays. Each tweet in JSON format can have over 150 attributes. In the case of my project, the attributes I will use are:

- *"id"*, which is the integer representation of the unique identifier of the tweet.
- *"text"'*, which contains the actual text of the tweet.
- "created_at", which is the exact date and time in UTC when the tweet was created.
- "Place", which contains the exact location that the tweet was created, if the tweet is geo tagged.

2.3 Pandas

Pandas is an open source software library written for the Python programming language for data manipulation and analysis. It is fast and efficient and one of the most preferred and widely used tools for data analysis. In particular, it can take data from files like CSV, text files, Microsoft Excel etc. and create a Dataframe object, where it gives the ability to perform many functionalities for the manipulation of data. Some of the functions are deleting and inserting data, reshaping datasets, joining and merging of datasets, and grouping and ordering of data (pandas [no date]). In addition, pandas can also be used for its methods of visualising and plotting diagrams on the screen. In the case of this project, pandas will be used for taking the data from a Microsoft Excel file (.xlsx) and for the utilisation of its methods for data preparation and analysis.

2.4 Sentiment Analysis

Sentiment Analysis is the process that combines Natural Language Processing (Hugging Face [no date]) and Machine Learning techniques to classify text into 'Neutral', 'Positive' or 'Negative' based on polarity scores, for the purposes of text analysis. It is rapidly becoming a crucial tool for monitoring and understanding sentiment in all types of data, as humans are expressing their feelings and thoughts more freely as time passes.

Sentiment analysis can be applied to any industry and can provide invaluable information to businesses and organisations. One use case of sentiment analysis is the use for the purposes of product analysis. Companies want to gather early feedback through reviews and comments on social media, forums, surveys and interaction with customer support for the continuous improvement of their product, even after its release. The ability to run sentiment analysis on the feedback data gives to the companies the opportunity to see why people have negative sentiment around their products and as a result, make modifications for the improvement based on them (AltexSoft 2018).

Another use case of sentiment analysis is social media monitoring. It enables businesses to gain insights into how their customers feel about specific topics, as well as discover crucial concerns in real-time (MonkeyLearn 2021). Despite that, it can be used for purposes of researching a specific topic and events in general, as in the case of this project. In this project, sentiment analysis will be applied to Twitter data that are related to electrical vehicles, in order to discover the sentiment and what the concerns of the people are about electrical vehicles.

The approaches that can be used for sentiment analysis are two, the Rule-based Sentiment Analysis and the Automated or Machine Learning Sentiment Analysis. The Rule-based is the more traditional way, which is based on manually created rules like lexicons. Lexicons are a list of positive and negative words. This technique counts the overall sentiment of the text based on the frequency of the words in the text. However, it does not consider the text as a whole and it is not a state-of-the-art method. On the other hand, the Automated Sentiment Analysis relies on machine learning classification algorithms. Classification models commonly use Logistic Regression, Linear Regression, Naive Baye, Support Vector Machines and Deep Learning. In this project, I will be using a roBERTa-base model, which is a Deep Learning Model. RoBERTa-base models are pre-trained on a large corpus of English data in a self-supervised method using Masked Language Modeling (MLM). In other words, they are trained only on raw texts, without the necessity of humans to label them, using an automatic process for the generation of inputs and labels from those texts (Hugging Face [no date]).

I have chosen this approach because Deep Learning methods and algorithms are state-of-the-art which outperform other techniques on most benchmarks. Additionally, they and are inspired by how the human brain is learning, and therefore can understand the context of the text better than the Rule-Based approach (Thematic [no date]).

2.4.1 Transformers

Transformers is a Python library which provides thousands of pre-trained models to perform tasks on different modalities and it is supported by the three most popular deep learning libraries such as 'Jax', 'PyTorch' and 'TensorFlow'. In the case of my project, this library is going to be used for text classification/sentiment analysis (Hugging Face [no date]). The

pipeline that I chose is the 'cardiffnlp/twitter-roberta-base-sentiment' which was developed by the Cardiff NLP research group. It uses a roBERTa-base model trained on approximately 58 million tweets and finetuned for sentiment analysis with the TweetEval benchmark. The model gets us an input text and returns the labels 0, 1 and 2 which mean 'Negative', 'Neutral' and 'Positive' respectively (Hugging Face [no date]).

2.5 Topic Modeling

Topic modeling is an *'unsupervised'* machine learning technique that can go over a collection of documents, find word and phrase patterns within them, and automatically cluster a set of keywords and related expressions that best represent the collection. It is called *'unsupervised'* because it does not depend upon training data or a previously defined list of tags classified by users. However, no one guarantees that this technique will give precise results.

Topic modeling can be applied in many real-world cases. For example, it can help in legal document searches to ensure that information is not missing. Due to the high volume of legal-related documents that need to be searched, it is easy to miss out on relevant facts. With topic modeling, effort and time can be saved as it helps in revealing sufficient information about the documents without even searching them all. Another real-world application of topic modeling is the content recommendation. New York Times, the American daily newspaper, in order to find the best match for the readers, it uses topic modeling to identify topics in articles, but also to identify topic preferences among readers (Rabindranath 2022). In this project, topic modeling will be applied to tweets, in order to identify the topics that people discuss on Twitter that are related to electrical vehicles. This will help in the discovery of people's concerns, but also the positive factors relating to electric vehicles.

There are several Topic Modeling methods such as, the Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Probabilistic Latent Semantic Analysis (PLSA) and Non-negative Matrix Factorization (NMF). In this project, I will be using the method of Latent Dirichlet Allocation (LDA) because it is one of the most commonly used.

2.5.1 LDA

Latent Dirichlet Allocation (LDA) was first introduced by Ng, and Jordan in 2003 and is one of the most prevalent methods in topic modeling. It is a probabilistic generative technique for modeling a corpus. The core concept is that the documents are represented as random mixes over latent topics, with each topic defined by a word distribution. LDA uses word probabilities to represent topics, which can usually give a good indication of what the topic is, by looking at the words with the highest probabilities in each topic (Jelodar et al. 2018).

2.5.2 Topic Coherence Score

The topic coherence score is a metric for evaluating an LDA topic model. It measures the degree of semantic similarity between high-scoring terms in a single topic. This metric helps in distinguishing between topics that are semantically interpretable topics and topics that are artifacts of statistical inference. The coherence measure that I will use in this project to evaluate the topic models is the 'C_v', which is one of the most commonly used. The 'C_v' measure is based on a sliding window, one-set segmentation of the top words, and an indirect confirmation

measure that employs normalised pointwise mutual information (NPMI) and the cosine similarity measure (Kapadia 2019).

2.5.3 Mallet (Machine learning for language toolkit)

Mallet is a java-based package for machine learning applications to text, such as statistical Natural Language Processing, topic modeling, document classification, information extraction, clustering and others (Mallet: MAchine Learning for LanguagE Toolkit. [no date]). The MALLET's toolkit for topic modeling contains sampling-based implementations of Latent Dirichlet Allocation, Pachinko Allocation, and Hierarchical LDA. I chose this package for the topic modeling because it has a more efficient implementation of LDA than Gensim's standard LDA implementation and it is known to be faster and provide better topic quality of topics (Prabhakaran 2018).

2.5.4 Natural Language Toolkit

The Natural Language Toolkit or NLTK is a python library that provides a series of text processing libraries for tokenization, classification, stemming and more, but also easy-to-use interfaces to corpora and lexical resources such as WordNet (NLTK :: Natural Language Toolkit. 2022). In this project, I will be using multiple libraries from the NLTK, which are going to help at the stage of the pre-processing of the tweets before passing them into the topic model. One of the tools I will use is the stop words, which include English words that do not add much meaning to a sentence and can safely be ignored without the risk of losing the meaning of the sentence. Stop words can be words such as "a", "the", "is", "are" etc. Another library available in NLTK that I will make use of is the Wordner Lemmetizer. Lemmatization is the process of converting a word into its lexical base form. For example, the word 'caring' will turn into 'care' and the word 'feet' to 'foot'. It prevents the topic model algorithm from counting "give," "gives," and "giving", for example, as three distinct word types.

2.5.5 Gensim

Gensim also known as "Generate Similar" is a popular open-source Python library for Natural Language Processing (NLP), used for the purposes of unsupervised Topic Modeling. It uses modern statistical machine learning and models to perform complex tasks, such as building document or word vectors, corpora, performing topic identification and more (Tutorialspoint [no date]). For this project, this library will be used for multiple tasks. Firstly, it will be used for the creation of the dictionary and corpus needed to build the LDA model using the object "corpora.Dictionary()". Also, this library will provide me with the wrapper for Latent Dirichlet Allocation (LDA), called "gensim.models.wrappers.LdaMallet", which allows LDA model estimation from training corpus and interference topic distribution on new and unseen documents as well (Tutorialspoint [no date]). Lastly, the library "models.coherencemodel" will be used to calculate the topic coherence of the topic models that will be created.

2.6 Data Analysis

2.6.1 Matplotlib

Matplotlib is a library for Python, which gives the ability to create static, animated, and interactive visualisations (Matplotlib 2021). In this project, I will use this library for the

visualisation of data, but also for the creation of graphs and figures in order to analyse the results of the sentiment analysis model.

2.6.2 Seaborn

Seaborn is another data visualisation library for Python, which provides a high-level interface for creating good-looking and informative statistical graphics (seaborn: statistical data visualization — seaborn 0.11.2 documentation. [no date]). This library will be used for the creation of visualisations and graphs for the purpose of analysing the results of the Topic Models and the data in general.

2.6.3 WordCloud

WordCloud is a Python library that gives the ability to create visualisations of word clouds. Word clouds is one of the data visualisation techniques that are most commonly used for the representation of text data, which indicates the importance or frequency of words based on their size. It will be used for the analysis of the tweets, but also for the representation of the topics that will be extracted from the topic models.

2.7 Related Work

Social media platforms are popular among researchers, due to the large amount of data and information that is readily available and free. Sentiment Analysis and Topic Modeling on social media, particularly on Twitter, is not a new concept, as many researchers use these methods already for their studies. For example, Boon-Itt and Skunkan (2020) conducted a study using Twitter data to discover what is the level of public awareness in terms of sentiments and emotions toward COVID-19 and what are the emergent topical themes and discourses regarding COVID-19. Another example is the research study conducted by Hidayatullah et al. (2018) which was focused on the discovery of the topic of the tweets about football news in Bahasa Indonesia.

Despite the fact that electric vehicles have been a trend for the last few years, there are not many research studies about the discourse of electric vehicles on social media using Natural language Processing techniques up to this date. In this section, I will review some of the key literature that I have found, not only for electric vehicles but for other topics as well. Specifically, I will discuss the methods and tools of collecting and analysing their data and also the results and conclusions from their studies.

Suresha and Kumar Tiwari (2021) conducted a study for the discovery of the perceptions and feelings of people regarding electric vehicles. In the study, the Twitter API was used for the collection of data using various hashtags '#' related to electric vehicles such as '#Electricfuture', '#Ev', '#Electricvehicle' etc. in order to find the tweets related to electric vehicles. The dataset consisted of 45000 tweets and had as columns the text, the hashtag, datetime and the user location of each tweet. For the pre-processing of the data, every uppercase was converted to its lowercase, URLs, user references, digits, stop words and repeated letters were removed, the text was tokenized and also POS (part-of-speech) tags were detected. For the method of Topic Modeling, the LDA via Gensim in conjunction with Mallet's implementation was used and for Sentiment Analysis, the library VADER was used, which is a lexicon-based and rule-based sentiment analysis tool. Apart from VADER, SONAR has also

been used for the identification of hateful and offensive tweets. The results of the study were that the top hashtag Twitter users tweeted while sharing tweets related to electrical vehicles was 'Tesla' and the most common location of users tweeting was Ekero Sweden. Also, it was found that 47.1% of tweets were positive, 42.4% were neutral, and 10.5% were negative as per VADER. Lastly, SONAR identified critical tweets which were hateful and offensive during the declaration of Tesla that they will not accept bitcoin for the purchase of electric vehicles. Even though in this article the LDA model has been used, the topics extracted have not been analysed which is something I will do in my project. Moreover, the topic model was not applied to positive and negative tweets separately, but to all the tweets regardless of sentiment. Compared to the method used in that article, in my project, I will apply topic modeling based on both positive and negative sentiment in order to find the topics in relation to each one.

Boon-Itt and Skunkan (2020) as mentioned previously, wanted to discover what is the level of public awareness in terms of sentiments and emotions toward COVID-19 and what are the emergent topical themes and discourses regarding COVID-19. In their study, the Twitter streaming API was used for data collection by specifying keywords and metadata such as language, source, data range, and location ending up with 107,990 tweets. For the pre-processing of the data, the tweet texts were subjected to a series of functions to remove URLs, emojis, special characters, retweets, hash symbols, and hyperlinks pointing to websites. Additionally, stop words in English were removed, the tweet text was converted into lower case and words were changed into their root form. For the sentiment analysis, the National Research Council (NRC) sentiment lexicon was used and basic emotions (e.g. anger, anticipation, disgust, fear etc.) were examined. For the Topic Modeling, the 'tidy()' method of the R package 'tidytext' was used. In the study, it was found that 22.12% of the tweets contained positive sentiments, while 77.88% contained negative sentiments. Moreover, some of the topics extracted from the topic model were 'The epidemic situation and confirmed cases of COVID-19' and 'Health concerns and fear as COVID19 is declared an emergency worldwide'.

The study of Kosaka (2022) does not have as a topic of interest the electric vehicles but the product Amazon Alexa. The author wanted to discover the sentiment of the people and the topics that are being discussed in the Amazon Alexa reviews. The dataset used in this study was obtained from Kaggle, consisting of 3000 Amazon customer reviews (input text), star ratings, date of review, variants and feedback of various amazon Alexa products like Alexa Echo, Echo dots, Alexa Firesticks etc. In this article, the pre-processing of the data included the creation of a new corpus with the words that occur more than ten times in the corpus of all the words of the dataset. Then, using this corpus, Gensim LDA topic modeling was applied in order to identify the most common topics. Sentiment analysis was performed using VADER. Sentiment ratings for different variations of Amazon Echo models were calculated and it was found that the positive was ten times higher than the negative. After that, the negative and positive sentiments were separated and using Count Vectorizer (TFIDF), the words that contributed to each sentiment were examined for the purposes of identifying the reasons why the positive and negative feedback was provided.

The final study I will review is the one of Nkuna (2020) which aimed to draw social media insights from the 2020 explosions in Lebanon. The tool TwitterScrapper was used for the collection of data using various trending hashtags '#' related to the Beirut Explosion, ending up with a dataset of 13,526 tweets. For the pre-processing of the data, all words were converted to lowercase, punctuation and non-alphabet characters were removed and cleaned in general with regex. After the cleaning, stopwords were removed, each sentence was tokenized to a

words list, bigrams and trigrams were created and the text was lemmatized. For the extraction of the topics, the article uses the method of LDA via Gensim so firstly, the corpus needed was created and then the topic model was applied. The main topics extracted were about the blast, praying and exposure. In this article, no sentiment analysis model was used.

Table 1 shows how the reviewed literature approached their studies.

Table 1- Approaches of other studies

Article	**Authors**	**Article Topic**	**Data Collection**	**Pre-processing of data**	**Sentiment analysis**	**Topic Modeling**	**Combination of Sentiment Analysis and Topic Modeling**
Topic Modeling and Sentiment Analysis of Electric Vehicles of Twitter Data	Suresha, H. and Kumar Tiwari, K.	Electric Vehicles	Twitter API	Convert into lowercase, removed URLs and user references, removed digits, removed stop words, removed repeated letters, tokenized and also detected POS tags.	VADER, SONAR	LDA Mallet model	No
Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study	Boon-Itt, S. and Skunkan, Y.	COVID-19	Twitter API	Removed URLs, emojis, special characters, retweets, hash symbols, and hyperlinks. Removed also stop words in English, converted into lower case and words were changed into their root form.	National Research Council (NRC) sentiment lexicon	LDA using tidytext (R)	No
Topic Modeling and Sentiment Analysis on Amazon Alexa Reviews	Kosaka, M.	Amazon Alexa	Kaggle	Creation of a new corpus with the words that occur more than ten times in the corpus of all the words of the dataset.	VADER	LDA Gensim model	No
What Twitter & Topic Mining reveal about the Beruit Explosion	Nkuna, J.	Beirut Explosion	TwitterScrapper	Lowercase, removed punctuation, removed non-alphabet characters, removed stopwords, tokenized each sentence to a words list, created bigrams and trigrams,lemmatized	No	LDA Gensim model	No

By reviewing the available literature, even though they analyse different topics, it has become clear that they have taken some common approaches which can be extracted. However, there is no literature available at the moment which followed the approach that I intend to adopt. The approaches followed in some of the reviewed literature and the approach followed in my project have some similarities. For example, I will be using the Twitter API to collect the data just like the first and second articles and the LDA Mallet model for topic modeling, just like in the first article. Additionally, the pre-processing stage of my project will be similar to the approaches of the first, second and third articles. However, some of the key differences in my approach is that I will be using a deep learning model (roBERTa-based model) for sentiment analysis instead of a lexicon rule-based model (VADER, SONAR, NRC Sentiment Lexicon) as articles one, two and three have done. Moreover, I will not perform sentiment analysis and topic modeling separately, but perform topic modeling on the positive and negative sentiments respectively, in order to identify the positive and negative topics.

2.8 Project Aim

The aim of this project is to analyse the discourse of people related to electric vehicles on social media and specifically on Twitter. To achieve this aim, a set of objectives must be accomplished including:

1. Retrieving the data related to Electrical Vehicles from Twitter and creating a dataset.
2. Pre-processing the tweets for the models.
3. Applying sentiment analysis to the tweets.
4. Applying topic modeling to the tweets.
5. Visualising and analysing the findings to gain insights.

3. Approach and Implementation

The following diagram shows the flow of the approach that I will take for the purposes of this project. Firstly, I will connect to the Twitter API in order to collect data related to electrical vehicles and create a dataset of tweets. Next, I will clean the noise from the tweets to be ready for the sentiment analysis model and then apply the model to classify the tweets into ‘Positive’, ‘Negative’ and ‘Neutral’. Then I need to pre-process the positive and negative tweets and create BOWs (Bag of Word) in order to be ready for topic modeling. After applying the topic modeling to the tweets, I will identify the interpretable topics extracted from the model and then conduct an analysis.

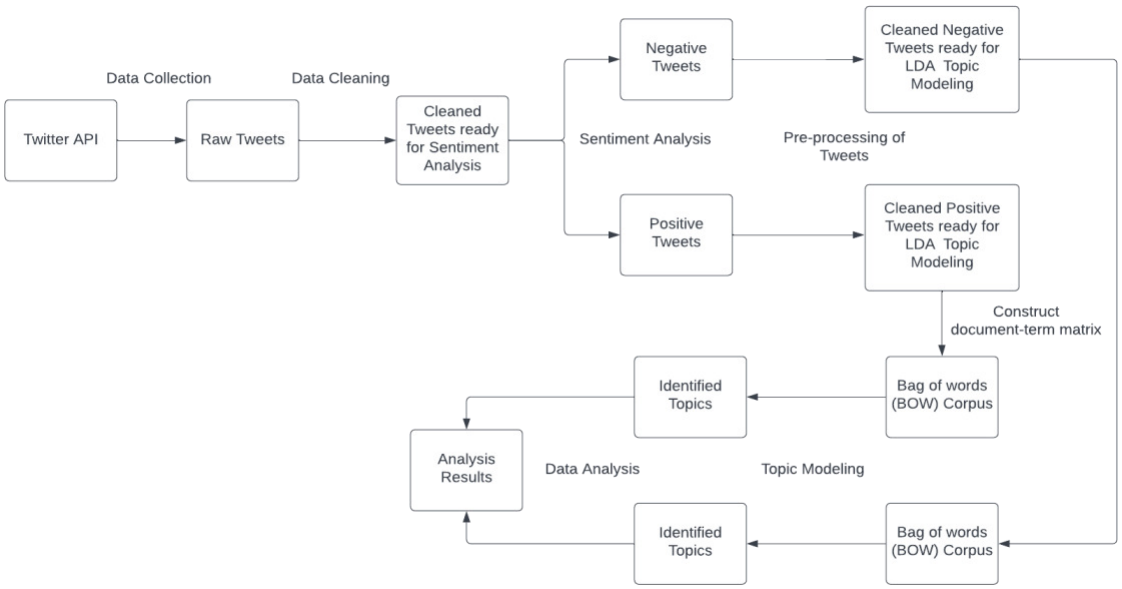

Figure 1 - Flow diagram of the approach

3.1 Creating a Database of Tweets

The goal of this project is to analyse the discourse that people have on Twitter regarding electric vehicles. There was no dataset available for this, so I needed to retrieve the data and create a dataset of tweets that have been tweeted globally and are related to electric vehicles.

3.1.1 Tweepy

The first step was to connect to the Twitter API using Tweepy and retrieve the tweets that are related to electric vehicles and save them in an Excel (xlsx) file. In order to do this, I firstly created a function for the client that connects to the Twitter API using the credentials that were provided to me when I created a Twitter developer account.

```
def getClient():
    client = tweepy.Client(bearer_token=config.bearer_token,
                           consumer_key=config.consumer_key,
                           consumer_secret=config.consumer_secret,
                           access_token=config.access_token,
                           access_token_secret=config.access_token_secret)
    return client
```

Figure 2 - Function for connection to the Twitter API using the credentials given

Then, I created a function that would allow me to stream tweets in real-time using the class "MyStreamListener". However, streaming the tweets was proven not to be very practicable as the program had to be running constantly in order to retrieve as many tweets as possible. Therefore, I created a function that would allow me to search Twitter's archive and retrieve tweets using custom filters in order to get the data I needed. I used the endpoint "Client.search_recent_tweets" which allowed me to retrieve tweets that match a search query, from the last seven days. The query I used for the tweets was to have either the keyword 'electric vehicles' or the hashtags '#electricvehicles', '#EVs' or '#EV', but also to be in English and not to be a retweet or a reply. I wanted to collect the initial tweets instead of retweets and

replies, because they would give a clearer image of the subject of the tweet. It is important to note that the query treats every filter as an 'AND' and not as 'OR' and that is why I used the operator 'OR' for the keyword and hashtags. Additionally, I was able to set the attribute of the tweets that I wanted to retrieve along with the tweets and those were the 'text', 'created_at' and 'geo', which represent the actual content of the tweet, the date and time that was posted and location. Even though I was able to get tweets only from the last seven days, I had the ability to select the period of time from which I wanted to retrieve the tweets. Due to the fact that I had Elevated access to Twitter's API and not an Academic Research one, I was only able to get a maximum of 100 tweets per request. Therefore, my approach was to retrieve 100 tweets every hour from 10:00 to 20:00 manually, therefore a total of 1000 tweets per day. I ended up collecting tweets between the 4th of February 2022 and the 27th of February 2022, overall 24 days and thus, creating a dataset of 18467 tweets with no distinct duplicates.

```
tweets = client.search_recent_tweets(query=query,
                                     tweet_fields=['text','created_at','geo'],
                                     place_fields=['full_name','geo'],
                                     expansions = 'geo.place_id',
                                     start_time='2022-03-02T19:00:00-00:00',
                                     end_time = '2022-03-02T20:00:00-00:00',
                                     max_results=100)
```

Figure 3 - Function which searches Twitter's archive

```
tweets = searchTweets('("electric vehicles" OR #electricvehicles OR #EVs OR #EV) lang:en -is:retweet -is:reply')
```

Figure 4 - Calling the function *'searchTweets'* with the query

Another thing that I noticed while writing this program, is that not all the tweets were geo-tagged. Therefore, I primarily had to check for it, and then if they were, I would save the location name associated with the tweet.

```
if not tweet.geo is None:
    if places[tweet.geo['place_id']]:
        place = places[tweet.geo['place_id']]
        location = place.full_name

date_created = (tweet.created_at.date())

excel_input_list.append(tweet_id)
excel_input_list.append(tweet_text)
excel_input_list.append(date_created)

if not tweet.geo is None:
    excel_input_list.append(location)
```

Figure 5 - If-else statement to see if the tweet is geo-tagged

All in all, the data included in the created Excel dataset were the text content of the tweets, the datetime, the location name if the tweets were geo-tagged, and the id of the tweets. For the representation and manipulation of the data, as well as for further analysis, I used pandas dataframe.

index	Tweet Id	Text	Date	Location
0	1.489750550019256e+18	Cutting-edge gallium nitride tech could help #EVs charge three times faster	2022-02-04 00:00:00	NaN
1	1.489750505442005e+18	Rattanindia Ent @ 57.00 /- #revolt #sustainme #rideforchange #rtnpower #electrical #bike #GoGreen #electricalbike #EVs #evswitch #rtnindia #rattanindia #rattanpower #drone #drones #fintech #ElectricVehicles #rattan #cocoblu #ecommerce @sustainme_in https://t.co/iaNy2nCNDM https://t.co/cTqZvmfNe3	2022-02-04 00:00:00	Bengaluru South, India
2	1.489750071025492e+18	Utilities' Carbon-Reduction Goals Will Have Little Impact On U.S. CO2 Emissions https://t.co/Pf4SJ7Gs6S #EV #VE #MOBILITY #NewMobility	2022-02-04 00:00:00	NaN
3	1.489749905488843e+18	Big Thanks to @CreativeMurdock! Great art work to keep us inspired in our investing journey! #EV #Lithium #Commodities #investing https://t.co/08AEubmZpb	2022-02-04 00:00:00	NaN
4	1.489749862782579e+18	Hurts me to post this (sorry fellow Tesla driver, but come on!!) but I feel like I have to….Don't be this type of driver 🤦 #ev #Rules #charging #Tesla #Model3 https://t.co/LZ31kLrenD	2022-02-04 00:00:00	NaN
5	1.489749618460144e+18	There are dozens of electric vehicles on the market today, but only one has been specifically engineered from the beginning with the discerning driving enthusiast in mind. https://t.co/nA0ddCPKNB	2022-02-04 00:00:00	NaN

Figure 6 - First five rows of the data

3.2 Cleaning the data

Before the tweets were processed, pre-processing was completed. The roBERTa-base sentiment analysis model that I would use is pre-trained on millions of tweets, so there is no need for further processing of the text other than cleaning it at this stage. To achieve this, I created a function that uses general practices of cleaning Twitter data with regular expressions (RegEx) and I have applied it to all the tweets data in the dataframe. The practices I have selected to apply were a combination of the practices that the reviewed literature used and contractions that I have noticed in the text. While I was looking for best practices for cleaning the text and going through articles, I noticed that people use similar methods, but not always the same. For example, methods I have used such as lowercasing, removing mentions, removing hashtags, removing URLs, removing 'RT' if the tweet is a retweet, removing the non-word characters and removing digits were taken from articles reviewed, and methods such as removing 'via' prior the hyperlinks, replacing the word 'amp' to 'and', removing unnecessary spaces, underscores and words containing numbers were methods that I have applied based on my observations in the text.

For the text to be ready for the sentiment analysis model the pre-processing was as follows:

- Lowercasing
- Removal of @mentions
- Removal of #hashtags
- Removal of the 'RT' which means Retweet in case the text is a retweet
- Removal of the 'via' word prior to hyperlinks
- Removal of the hyperlinks
- Removal of any non-word characters
- Replacement of the word 'amp' with the word 'and'
- Removal of the words containing numbers
- Removal of digits that are not concatenated to words
- Removal of spaces at the end of the line
- Removal of spaces at the beginning of the line
- Removal of underscores
- Replacement of all multiple white spaces with a single white space

However, to avoid any word sense disambiguation in text, before cleaning the text, I have called another function that converts the contracted words in the text into standard lexicons. For example, it converts the "don't" to "do not" and "can't" to "can not". I noticed that in the text there are two types of apostrophes, the " ' " and " ' " thus I have covered both cases.

```
#Replacing apostrophes/short words
def decontracted(phrase):
  # with the apostroph -> ’
  phrase = re.sub(r"won\’t", "will not", phrase)
  phrase = re.sub(r"can\’t", "can not", phrase)
  phrase = re.sub(r"n\’t", " not", phrase)
  phrase = re.sub(r"\’re", " are", phrase)
  phrase = re.sub(r"\’s", " is", phrase)
  phrase = re.sub(r"\’d", " would", phrase)
  phrase = re.sub(r"\’ll", " will", phrase)
  phrase = re.sub(r"\’t", " not", phrase)
  phrase = re.sub(r"\’ve", " have", phrase)
  phrase = re.sub(r"\’m", " am", phrase)

  # with the apostroph -> '
  phrase = re.sub(r"won\'t", "will not", phrase)
  phrase = re.sub(r"can\'t", "can not", phrase)
  phrase = re.sub(r"n\'t", " not", phrase)
  phrase = re.sub(r"\'re", " are", phrase)
  phrase = re.sub(r"\'s", " is", phrase)
  phrase = re.sub(r"\'d", " would", phrase)
  phrase = re.sub(r"\'ll", " will", phrase)
  phrase = re.sub(r"\'t", " not", phrase)
  phrase = re.sub(r"\'ve", " have", phrase)
  phrase = re.sub(r"\'m", " am", phrase)
  return phrase
```

Figure 7 - Converting contracted words in the text to standard lexicons

```
def cleanText(text):
  text = text.lower() # Turn the whole text to lowercase
  text = decontracted(text) # Calling the above function
  # r tells python that is raw stream
  text = re.sub(r'@[a-z0-9]+', '', text) # Removes @mentions
  text = re.sub(r'#[a-z0-9]+', '', text) # Removes the hashtags
  text = re.sub(r'rt[\s]+', '', text) # Removes the RT
  text = re.sub(r'via','', text) # Removes via word prior the hyper link
  text = re.sub(r'https?:\/\/\S+', '', text) # Removes the hyper link
  text = re.sub(r'\W+',' ', text) # Removes any non word characters
  text = re.sub(r'amp','and', text) # Replaces 'amp' with 'and'
  text = re.sub(r'\w*\d\w*', '', text) # Removes words containing numbers
  text = re.sub(r' \d+', '', text) # Removes the digits that are not concatinated in words
  text = re.sub(r'\s$','', text) # Removes spaces at the end of line
  text = re.sub(r'^\s','', text) # Removes spaces at the beginning of line
  text = re.sub(r'_','',text) # Removes underscores
  text = re.sub(r'\s+',' ',text) # Replace all multiple white spaces with single white space

  return text
```

Figure 8 - Process of cleaning the text

When the data were cleaned, I removed the duplicates again because there were identical tweets with different URLs, leaving a dataset of a total of 15603 tweets. In figure 9, it can be seen that after the application of the functions, the text has been cleaned and has been added to a new column called “Cleaned Text” in the dataframe.

Tweet Id	Text	Date	Location	Cleaned Text
1.489751e+18	Cutting-edge gallium nitride tech could help #...	2022-02-04	NaN	cutting edge gallium nitride tech could help c...
1.489751e+18	Rattanindia Ent @ 57.00 /-\n\n #revolt #sustai...	2022-02-04	Bengaluru South, India	rattanindia ent in
1.489750e+18	Utilities' Carbon-Reduction Goals Will Have Li...	2022-02-04	NaN	utilities carbon reduction goals will have lit...
1.489750e+18	Big Thanks to @CreativeMurdock! \n\nGreat art ...	2022-02-04	NaN	big thanks to great awork to keep us inspired ...
1.489750e+18	Hurts me to post this (sorry fellow Tesla driv...	2022-02-04	NaN	hurts me to post this sorry fellow tesla drive...
...	...	...	...	...
1.489728e+18	Artist Michael Doyle brings us to the edge of ...	2022-02-04	NaN	artist michael doyle brings us to the edge of ...
1.489727e+18	If Detroit can build an experimental road that...	2022-02-04	NaN	if detroit can build an experimental road that...
1.489726e+18	France's January plugin #electricvehicles shar...	2022-02-04	NaN	france is january plugin share at with full el...
1.489726e+18	The first week in February has been full of ex...	2022-02-04	NaN	the first week in february has been full of ex...
1.489726e+18	Sweden in January: The country has now flipped...	2022-02-04	NaN	sweden in january the country has now flipped ...

Figure 9 - Sample of cleaned tweets

3.3 Sentiment Analysis

3.3.1 Classifying Tweets to 'Positive', 'Negative' and 'Neutral'

Once the tweets in the dataset were cleaned from the 'noise', the next step was to define the sentiment of each tweet. In the beginning, I used a lexicon-based sentiment analysis model, the 'TextBlob', as I had seen many articles using it. However, the results of the topic models were not very interpretable, as this sentiment analysis model did not accurately classify the tweets to a sentiment. As a result, the topic models were extracting the same topics from both positive and negative tweets. Therefore, I used a roBERTa-based model developed by the Cardiff NLP research group.

I first set the pipeline with the model 'cardiffnlp/twitter-roberta-base-sentiment' using the library 'Transformers' by assigning it to the variable with the name 'sentiment_analysis_model'.

```
sentiment_analysis_model = pipeline(model="cardiffnlp/twitter-roberta-base-sentiment")
```

Figure 10 - Initialising the sentiment analysis pipeline

There are three steps involved when text is passed to the pipeline. Firstly, the text is pre-processed into an understandable format for the model, then the pre-processed input is passed to the model and lastly, the predictions of the model are post-processed so that it can be human-readable (Hugging Face [no date]).

After the setup of the pipeline, I created a function that gets as input the cleaned text from the dataframe and returns the prediction of the pipeline. In the case of the model used, the predictions that are returned are the string label 'LABEL_2' if the tweet is positive, 'LABEL_0' if the tweet is negative and 'LABEL_1' if it is neutral. Hence, in order for the output to be more understandable, I implemented an if-else statement to return the strings 'Positive', 'Negative' and 'Neutral' instead.

```
#Labels: 0 -> Negative; 1 -> Neutral; 2 -> Positive
def sentimentAnalysis(text):
  sentimentAnalysis.counter += 1
  print(sentimentAnalysis.counter)
  if sentiment_analysis_model(text)[0]['label'] == 'LABEL_2':
    return 'Positive'
  elif sentiment_analysis_model(text)[0]['label'] == 'LABEL_0':
    return 'Negative'
  else:
    return 'Neutral'
sentimentAnalysis.counter = 0
df['Sentiment Analysis'] = df['Cleaned Text'].apply(sentimentAnalysis)
```

Figure 11 - Sentiment Analysis classification function

Cleaned Text	Sentiment Analysis
rattanindia ent in	Neutral
big thanks to great awork to keep us inspired ...	Positive
hurts me to post this sorry fellow tesla drive...	Negative
new analysis from atlas public policy finds el...	Positive
it is tax season did you know that there are a...	Positive

Figure 12 - Sample of classified tweets to Neutral, Positive, Negative

In order to proceed to the topic modeling process, I first needed to be sure that the outcome of the sentiment analysis model that I had used made sense. Therefore, I explored the positive and negative tweets and they seemed to be quite reasonable.

Some of the positive tweets were:

- *"working towards a greener future qml"*
- *"this all electric tractor helps farms go green"*
- *"this is a great step forward"*
- *"electric vehicles sales to go up in next years"*
- *"but electric vehicles are so much safer"*

Some of the negative tweets were:

- *"motor mouth why i am not completely in love with electric vehicles yet"*
- *"i bet all those that brought electric cars are going to regret it come april"*
- *"i do not know much about electr*ic cars but with how much does it cost to charge one *will it become too expensive to own one"*
- *"so the government push electric vehicles down our throats and allow a increase in electric prices they really do take us for mugs"*
- *"i am not enthusiastic about any electric vehicles not even tesla"*

3.4 Pre-processing of the tweets for LDA Topic Modeling

After the classification of the tweets to 'Positive', 'Neutral' and 'Negative', pre-processing of the tweets was required in order to be used in the topic model. I already cleaned the text at this point for the purposes of the sentiment analysis model. Hence, the next step was to pre-process

the tweets using various Natural Language Processing methods in order to be ready to be passed into the topic model. I have used general practices of pre-processing such as tokenization, removal of the stopwords, lemmatization, and creation of bigrams and trigrams.

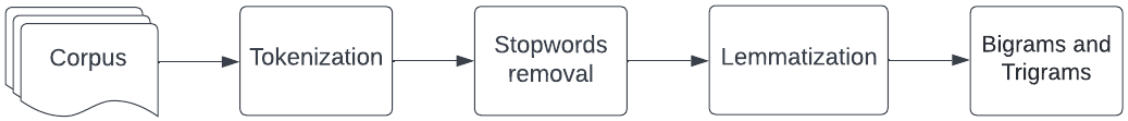

Figure 13 - Pre-processing steps of the corpus

3.4.1 Tokenization of the cleaned text

The first stage of the pre-processing was the tokenization of the text. Tokenizing is splitting some string or sentence into a list of words. This stage was necessary in order to proceed to the next stages of pre-processing. I have applied a function on each row of the dataframe, which uses regex to split the text when it sees one or more non-word characters such as white spaces.

```
def tokenize(text):
    tokens = re.split(r'\W+', text)
    return tokens

df['Cleaned Text Tokenized'] = df['Cleaned Text'].apply(lambda x: tokenize(x.lower()))
```

Figure 14 - Tokenize function

Cleaned Text	Sentiment Analysis	Cleaned Text Tokenized
rattanindia ent in	Neutral	[rattanindia, ent, in]
big thanks to great awork to keep us inspired ...	Positive	[big, thanks, to, great, awork, to, keep, us, ...
hurts me to post this sorry fellow tesla drive...	Negative	[hurts, me, to, post, this, sorry, fellow, tes...
new analysis from atlas public policy finds el...	Positive	[new, analysis, from, atlas, public, policy, f...
it is tax season did you know that there are a...	Positive	[it, is, tax, season, did, you, know, that, th...

Figure 15 - Sample of tokenized text

3.4.2 Remove stopwords from the tokenized cleaned text

The next stage of pre-processing was the removal of the stopwords from the text. Stopwords are a set of words that are commonly used (such as, “a”, “in”, “an”, “the”) that do not provide much meaningful context on their own. The function that I created uses the NLTK’s list of stopwords for their removal from the text. It basically gets the tokenized lists of the tweets from

the dataframe and removes the stopwords from them, if there are any. The function was applied to the tokenized text in the dataframe and saved in a new column 'Cleaned Text NoStop'.

```
stopword = nltk.corpus.stopwords.words('english')

def remove_stopwords(tokenized_list):
    text = [word for word in tokenized_list if word not in stopword]
    return text

df['Cleaned Text NoStop'] = df['Cleaned Text Tokenized'].apply(lambda x: remove_stopwords(x))
```

Figure 16 - Removing stopwords function

Cleaned Text	Sentiment Analysis	Cleaned Text Tokenized	Cleaned Text NoStop
rattanindia ent in	Neutral	[rattanindia, ent, in]	[rattanindia, ent]
big thanks to great awork to keep us inspired ...	Positive	[big, thanks, to, great, awork, to, keep, us, ...	[big, thanks, great, awork, keep, us, inspired...
hurts me to post this sorry fellow tesla drive...	Negative	[hurts, me, to, post, this, sorry, fellow, tes...	[hurts, post, sorry, fellow, tesla, driver, co...
new analysis from atlas public policy finds el...	Positive	[new, analysis, from, atlas, public, policy, f...	[new, analysis, atlas, public, policy, finds, ...
it is tax season did you know that there are a...	Positive	[it, is, tax, season, did, you, know, that, th...	[tax, season, know, sorts, state, federal, reb...

Figure 17 - Sample of tweets without stopwords

3.4.3 Lemmatization

What came next, was to condense derived words in the text into their base forms. This task could be achieved by two methods, lemmatization and stemming. For my project, I have chosen to use the method of lemmatizing, because it is typically more accurate than stemming, since it relies on a language's full vocabulary to apply a morphological analysis to words (Beri 2020). On the contrary, Stemming is typically faster as it simply chops off the end of a word using heuristics, but it does not have any understanding of the context in which a word is used (Mal 2021). For instance, lemmatization would correctly identify the base form of the word 'caring' to 'care' compared to stemming that would return just 'car'. Additionally, according to an experiment run by Bitext (2017), lemmatization improves the results achieved while using the Topic Modeling algorithm.

To do this, I have created the lemmatizing() function, where I used the WordNet Lemmatizer which is using WordNet, a large lexical database for English. It gets as input the tokenized text and returns the text with the words in their base form. The function calls the function lemmatize() of the instance WordNetLemmatizer() and lemmatizes every single word in the tokenized text. However, in order for the lemmatizer to work correctly, I have provided the correct POS (Part of Speech) tag as a second argument to the lemmatize() function. POS tagging is the process of marking up words in text based on their definition and context. For example, the POS tag of the word 'ask' would be 'VB' because it is a verb (Johnson 2022). It is not feasible to provide manually the correct POS for every word in a large set of texts, therefore, I have used another function called getWordnetPos() that finds out the correct POS of the word

using the nltk.pos_tag() (Prabhakaran 2018). The nltk.pos_tag() returns a tuple with the POS tag, so mapping of the POS tag to the format wordnet lemmatizer would accept was needed. The function lemmatizing() was applied to the tokenized text with no stopwords in the dataframe and saved in a new column 'Clean Text NoStop Lemmatized'.

```
def getWordnetPos(word):
    posTag = nltk.pos_tag([word])[0][1][0].upper()
    tagDict = {"J": wordnet.ADJ,
               "N": wordnet.NOUN,
               "V": wordnet.VERB,
               "R": wordnet.ADV}

    return tagDict.get(posTag, wordnet.NOUN)
```

Figure 18 - POS tags function

```
wn = nltk.WordNetLemmatizer()
def lemmatizing(tokenized_text):
    text = [wn.lemmatize(word, getWordnetPos(word)) for word in tokenized_text if word]
    return text

df['Clean Text NoStop Lemmatized'] = df['Cleaned Text NoStop'].apply(lambda x: lemmatizing(x))

df.head(8)
```

Figure 19 - Lemmatizer function

Cleaned Text	Sentiment Analysis	Cleaned Text Tokenized	Cleaned Text NoStop	Clean Text NoStop Lemmatized
rattanindia ent in	Neutral	[rattanindia, ent, in]	[rattanindia, ent]	[rattanindia, ent]
big thanks to great awork to keep us inspired ...	Positive	[big, thanks, to, great, awork, to, keep, us, ...	[big, thanks, great, awork, keep, us, inspired...	[big, thanks, great, awork, keep, u, inspire, ...
urts me to post this sorry fellow tesla drive...	Negative	[hurts, me, to, post, this, sorry, fellow, tes...	[hurts, post, sorry, fellow, tesla, driver, co...	[hurt, post, sorry, fellow, tesla, driver, com...
new analysis from atlas public policy finds el...	Positive	[new, analysis, from, atlas, public, policy, f...	[new, analysis, atlas, public, policy, finds, ...	[new, analysis, atlas, public, policy, find, e...
it is tax season did you know that there are a...	Positive	[it, is, tax, season, did, you, know, that, th...	[tax, season, know, sorts, state, federal, reb...	[tax, season, know, sort, state, federal, reba...
from bills aimed at promoting electric vehicle...	Positive	[from, bills, aimed, at, promoting, electric, ...	[bills, aimed, promoting, electric, vehicles, ...	[bill, aim, promote, electric, vehicle, carbon...
inergency the ambition loop in motion for elec...	Neutral	[inergency, the, ambition, loop, in, motion, f...	[inergency, ambition, loop, motion, electric, ...	[inergency, ambition, loop, motion, electric, ...
detroit we will build one mile of electrified ...	Negative	[detroit, we, will, build, one, mile, of, elec...	[detroit, build, one, mile, electrified, road,...	[detroit, build, one, mile, electrify, road, c...

Figure 20 - Sample of the lemmatized text

3.4.4 Bigrams and Trigrams

The final stage of pre-processing of the text was to create bigrams and trigrams. Bigrams are 2 words frequently occurring in the text and trigrams are 3. I have created and used a function that creates both bigrams and trigrams in the tokenized text of the positive and negative tweets separately, using Gensim's model '*models.phrases'*. I have set the argument 'min_count' to be 5 to ignore all the bigrams with a total collected count lower than this value, and the argument 'threshold' to be 100 which represents the score threshold for forming the phrases.

```
# Function that returns the tokenze)list with bigrams and trigrams
def bigram_trigram(tokenize_list):

    bigram_phrases = gensim.models.Phrases(tokenize_list, min_count=5, threshold=100)
    trigram_phrases = gensim.models.Phrases(bigram_phrases[tokenize_list], threshold=100)

    bigram = gensim.models.phrases.Phraser(bigram_phrases)
    trigram = gensim.models.phrases.Phraser(trigram_phrases)

    data_bigrams = [bigram[doc] for doc in tokenize_list]
    data_bigrams_trigrams = [trigram[bigram[doc]] for doc in data_bigrams]

    return data_bigrams_trigrams
```

Figure 21 - Bigrams and Trigram creation function

Some of the resulting bigrams and trigrams were ‘San_francisco’, 'general_motor', 'air_quality', 'porsche_taycan', ’fuel_cell’, 'fossil_fuel', 'long_term', 'chief_executive', 'low_carbon', 'climate_change', 'renewable_energy’ etc.

3.5 LDA Topic Modeling

After the pre-processing stage of the tweets, the following task was to perform a topic model in order to identify topics in the positive and negative tweets. The topic model approach that I have chosen to take was the one of the Latent Dirichlet Allocation (LDA). LDA uses a generative probabilistic model and Dirichlet distributions to discover topics that are latent in a corpus (set of documents). If K topics characterise a set of documents, the mix of topics in each document can be described by a K-nomial distribution, which is a type of multinomial distribution. LDA uses two Dirichlet distributions in its algorithm, one over the K-nomial distributions of topic mixes and one over the words in each topic.

First of all, LDA needs the number of topics K which is believed to better describe the set of documents being analysed. K is an important parameter of LDA so if there is uncertainty about the number of topics, the trial-and-error approach can be used. Besides K, two other parameters in LDA are the Alpha (α) and Eta (η) which are associated with the two Dirichlet distributions and act as ‘concertation’ parameters (Rabindranath 2022). Firstly, Alpha is the document-topic density at which the higher the values, the more the topics which will be composed in the documents, and for lower values the documents will have fewer topics. Secondly, Eta is the topic-word density at which the higher its value is, the greater the number of words which will be in a given topic and the lower its value, the fewer the words which will be in the topics. The choice of the Alpha and Eta parameters can play an important role in the topic modeling algorithm. However, popular implementations of LDA set default values for these parameters and can be changed manually (Seth 2020).

An important thing to note is that LDA does not tell exactly what the topics are, other than showing the distribution of words contained within them. Therefore, in order to give names to the topics and choose the ones that make some sense, interpretation and some knowledge of the subject being analysed are needed.

LDA is essentially an iterative process of assigning topics to each word in each document in the set, in order to generate the distribution of topics for each document. The algorithm of LDA has 4 steps and can be described as follows (Rabindranath 2022):

Step 1) Initialization of the model:

- Assignment of a topic to each word in each document at random
- Computation of frequency distribution of the topics in each document (topic frequency)
- Computation of frequency distribution of words in each topic (word frequency)

Step 2) Update of the topic assigned on a single word in a single document:

- Choice of word in a document
    - Un-assignment of the assigned topic of the word
    - Re-assignment of a topic to the word given all other topic assignments for all other words in all documents by taking into account:
        - The number of times the document uses each topic, measured by the topic frequency calculated during step 1 and a Dirichlet-generated multinomial distribution over topics for each document
        - The number of times each topic uses the word, measured by the word frequency calculated in step 1 and a Dirichlet-generated multinomial distribution over words for each topic
    - Getting the conditional probability that the word takes on each topic by multiplying the number of times the document uses each topic and the number of times each topic uses the word
    - Re-assignment of the word to the topic with the largest conditional probability

Step 3) Update of the topic already assigned on a single word in a single document:

Repetition of Step 2 for all words in all documents.

Step 4) Iteration

3.5.1 Construct document-term matrix

The main inputs to the LDA topic model algorithm are a document word matrix (corpus) and a dictionary. The "corpora.Dictionary()" object of Gensim has given me the ability to create two dictionaries for both positive and negative tweets, with the mapping of all words and their unique id from the lists of the tokenized pre-processed tweets. These dictionaries have then been used for the creation of the Bag of Words (BOW) corpora, which contained a list of lists of tuples containing word token ids, along with their frequency count in each tweet. (Tutorialspoint [no date]).

```
positive_id2word = corpora.Dictionary(positive_bigrams_trigrams)
negative_id2word = corpora.Dictionary(negative_bigrams_trigrams)
```

Figure 22 - Creation of dictionaries

```
Dictionary(8522 unique tokens: ['awork', 'big', 'great', 'inspire', 'invest']...)
Dictionary(4549 unique tokens: ['come', 'driver', 'feel', 'fellow', 'hurt']...)
```

Figure 23 - Results of dictionaries

```
positive_corpus = [positive_id2word.doc2bow(text) for text in positive_texts]
negative_corpus = [negative_id2word.doc2bow(text) for text in negative_texts]
```

Figure 24 - Creation of BOW

For example, as can be seen in figure 25, the word id 0 occurs once in the document. Likewise, word id 1 occurs twice and so on.

```
[[(0, 1), (1, 2), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1)],
```

Figure 25 - Sample of BOW

3.5.2 Applying LDA Mallet Model

Before I started using the Mallet with Gensim for LDA, I downloaded the mallet-2.0.8.zip package on google colab and unzipped it. Once it was downloaded, I set the path of the file to be the content of the mallet in order to use it when creating the models. I have applied the LDA Mallet model on both positive and negative tweets respectively, in order to identify the topics from each sentiment.

I have used a function that builds many LDA topic models with different values of numbers of topics (K) as an input. The reason for doing this, was to find the optimum number of topics in order to get the best possible results from the model. Choosing a number of K topics which denotes the end of a rapid increase in topic coherence, usually gives interpretable and meaningful topics. However, picking an even higher value can sometimes provide more granular sub-topics (Prabhakaran 2018). The parameters of the function are the dictionary (id2word), the corpus, the numbers of the range of the for-loop that generates the models (limit, start, step) and also the texts which is the list of the pre-processed text. The purpose of this function is to iterate through a loop using the input range numbers, to build models with a different topic number K as an input, and then append the models in a list. At the same time, it computes the coherence score of each model. All the other parameters of the LDA Mallet model such as alpha, workers, prefix, optimize_interval, iterations, topic_threshold and random_seed were set to be the defaults.

```
def compute_coherence_values(dictionary, corpus, texts, limit, start, step):
    """
    Compute c_v coherence for various number of topics

    Parameters:
    ----------
    dictionary : Gensim dictionary
    corpus : Gensim corpus
    texts : List of input texts
    limit : Max num of topics

    Returns:
    -------
    model_list : List of LDA topic models
    coherence_values : Coherence values corresponding to the LDA model with respective number of topics
    """
    coherence_values = []
    model_list = []
    for num_topics in range(start, limit, step):
        model = gensim.models.wrappers.LdaMallet(mallet_path, corpus=corpus, num_topics=num_topics, id2word=dictionary)
        model_list.append(model)
        coherencemodel = CoherenceModel(model=model, texts=texts, dictionary=dictionary, coherence='c_v')
        coherence_values.append(coherencemodel.get_coherence())

    return model_list, coherence_values
```

Figure 26 - LDA Mallet Models creation and Coherence Score function

For both positive and negative tweets, the inputs of the models were the previously created dictionary, corpus, texts and also the numbers of the intervals for the topics to be from 2 to 40 incrementing by 6 (2, 8, 14, 20, 26, 32, 40).

```
positive_model_list, positive_coherence_values = compute_coherence_values(dictionary=positive_id2word, corpus=positive_corpus, texts=positive_texts, start=2, limit=40, step=6)
```

Figure 27 - Creation of Positive LDA Mallet Models with their coherence score

```
negative_model_list, negative_coherence_values = compute_coherence_values(dictionary=negative_id2word, corpus=negative_corpus, texts=negative_texts, start=2, limit=40, step=6)
```

Figure 28 - Creation of Negative LDA Mallet Models with their coherence score

After the creation of the models, I plotted the coherence scores of the models in order to choose the model with the number of topics with a relatively good topic coherence that would give the best possible topics. In the case of the positive tweets topic model, I have chosen the number K=14 topics, because in the graph it denoted the end of a rapid increase where it usually offers meaningful and interpretable topics (Prabhakaran 2018). That model gave a relatively good coherence score (0.3658) compared to the other models with different numbers of topics. The topics were contributed by the top 10 keywords with the highest probability distribution, so I qualitatively inspected and discarded topics that were noisy and irrelevant.

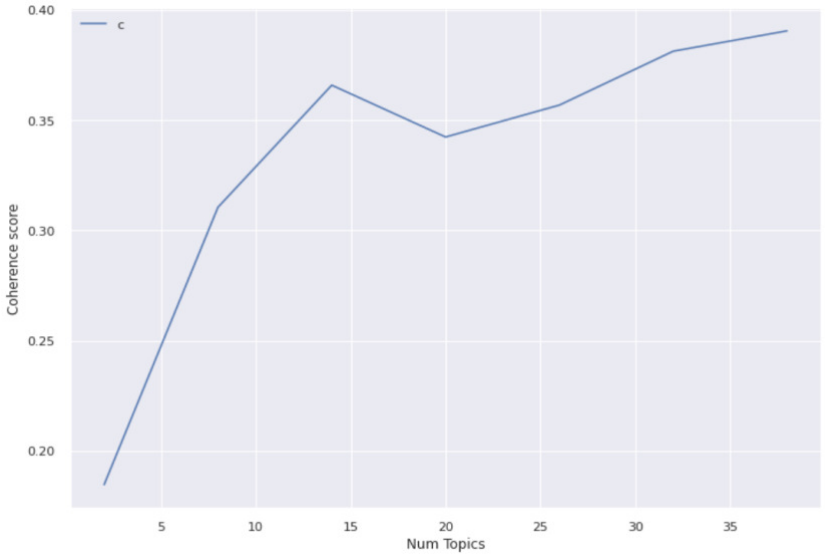

Figure 29 - Graph showing the Coherence Scores of the Positive topic models

```
Num Topics = 2  has Coherence Value of 0.1846
Num Topics = 8  has Coherence Value of 0.3104
Num Topics = 14  has Coherence Value of 0.3658
Num Topics = 20  has Coherence Value of 0.3423
Num Topics = 26  has Coherence Value of 0.3568
Num Topics = 32  has Coherence Value of 0.3812
Num Topics = 38  has Coherence Value of 0.3904
```

Figure 30 - Coherences Scores of the Positive topic models for each number of topics

In the case of the negative tweets topic model, I have chosen the number k=26 topics, for the same reason that I have chosen to do so for the purposes of the positive tweets topic model (Prabhakaran 2018). The coherence score of that model was 0.5089. The topics were contributed by the top 10 keywords with the highest probability distribution, so once again I qualitatively inspected and discarded topics that were noisy and irrelevant.

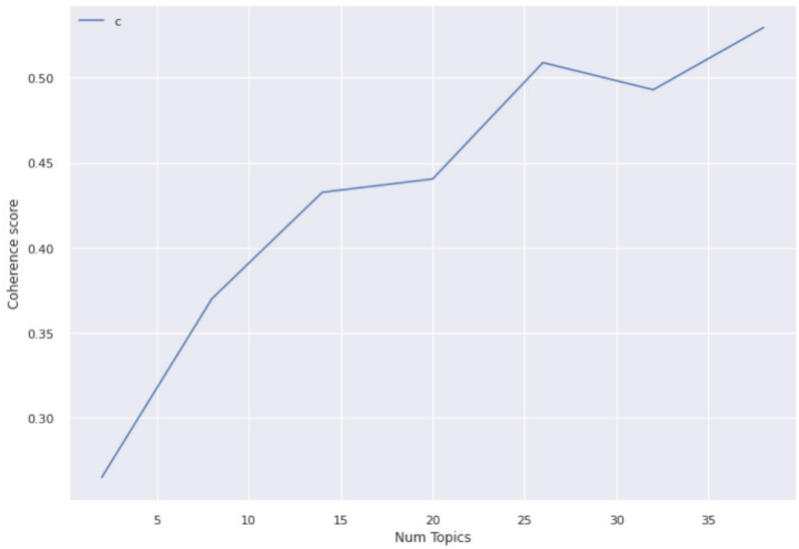

Figure 31 – Graph showing the Coherence Scores of the Negative topic models

```
Num Topics = 2  has Coherence Value of 0.2648
Num Topics = 8  has Coherence Value of 0.3701
Num Topics = 14  has Coherence Value of 0.4326
Num Topics = 20  has Coherence Value of 0.4404
Num Topics = 26  has Coherence Value of 0.5089
Num Topics = 32  has Coherence Value of 0.493
Num Topics = 38  has Coherence Value of 0.5296
```

Figure 32 - Coherences Scores of the Negative topic models for each number of topics

4. Results

After the collection and the processing of the dataset, the final stage was to analyse the tweets and gain insights from the Sentiment Analysis and Topic Modeling techniques. The tweets were analysed and visualised using various graphs and pie charts but also using word clouds. Firstly, I analysed the data before the pre-processing stage and their input to the models, in order to get some statistics and understand the raw data better.

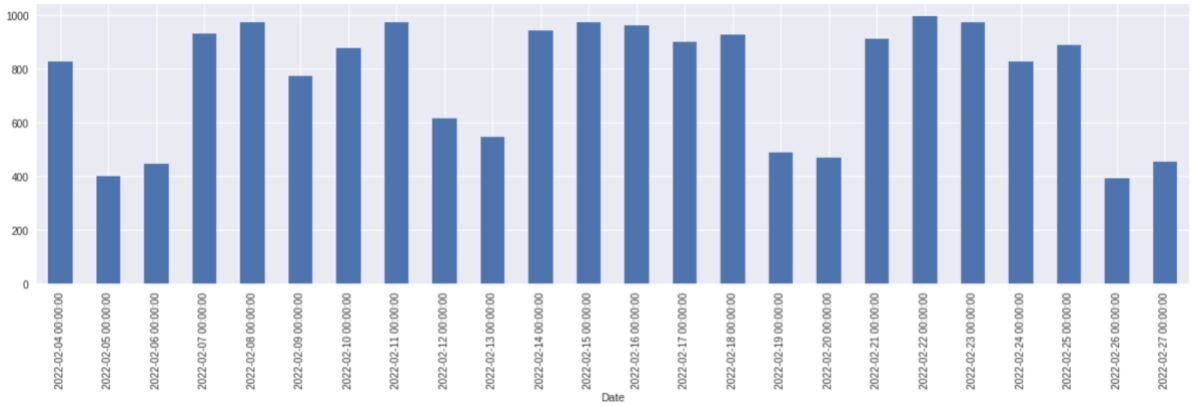

Figure 33 - Bar plot of the numbers of tweets collected daily

The total number of tweets in the dataset before the cleaning stage was 18467. As it can be seen in figure 33, the number of tweets daily was varying between approximately 400 to 1000 tweets. This is because, even though 1000 tweets were retrieved each day, I was removing the distinct duplicates using Microsoft Excel.

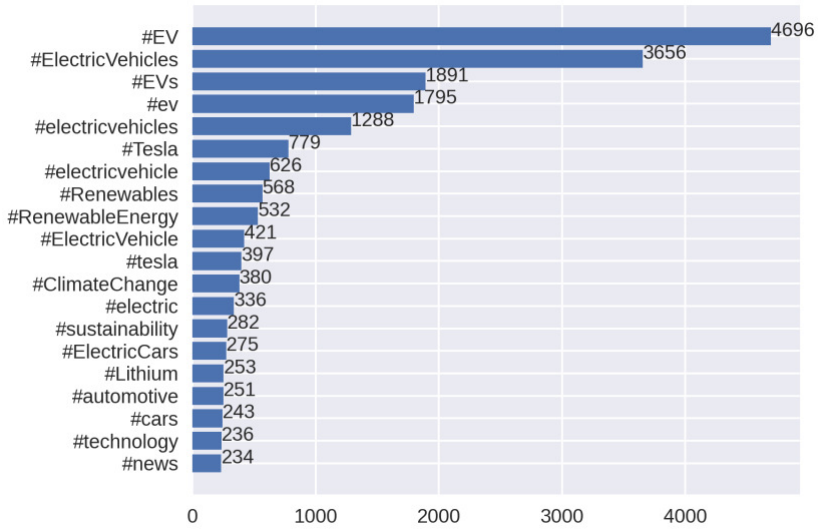

Figure 34 - Bar plot of the 20 most used hashtags in the tweets

In the bar chart in figure 34, the 20 most used hashtags in the tweets are presented. The most posted hashtag was the #EV which was present in 4696 tweets. There were also other hashtags for electric vehicles in different formats e.g. #ev, #ElectricVehicles, #EVs etc. Moreover, some other topics that can be related to electric vehicles appeared in the chart including Tesla, renewable energy, Lithium, climate change, sustainability, technology, automotive and charging.

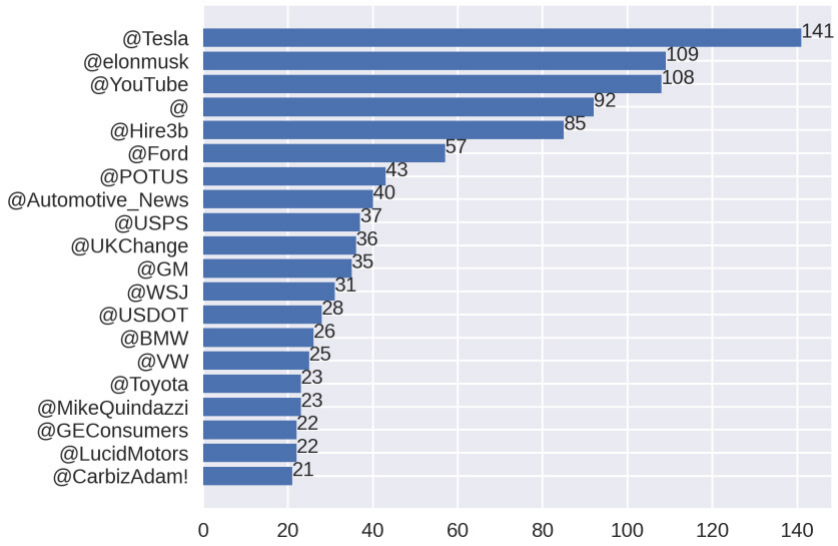

Figure 35 - Bar plot of the 20 most used mentions in the tweets

In the bar chart in figure 35, the 20 most used mentions in the tweets are presented. It appears that the top 2 were the company Tesla and its owner Elon Mask, with 141 and 109 mentions respectively. Also, some other companies and brands were mentioned, such as Youtube, Hire3b, Ford, BMW etc.

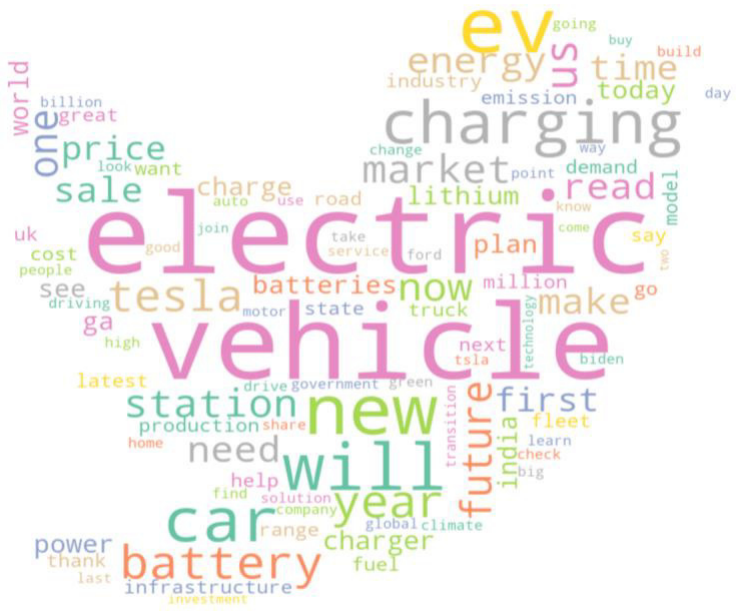

Figure 36 - Word cloud of the top 100 most common words in all the tweets

Figure 36 shows a word cloud of the top 100 most common words found in all the tweets. As I collected tweets that had the keywords of 'electric vehicles', it was expected that the most frequent words in the tweets would be 'electric' and 'vehicles'. Moreover, other frequent words that seemed to be interesting were 'charging', 'station', 'battery', 'tesla', 'charger', 'market', 'price' and 'sale'.

```
Sentiment Analysis
Negative    1505
Neutral     9000
Positive    5098
dtype: int64
```

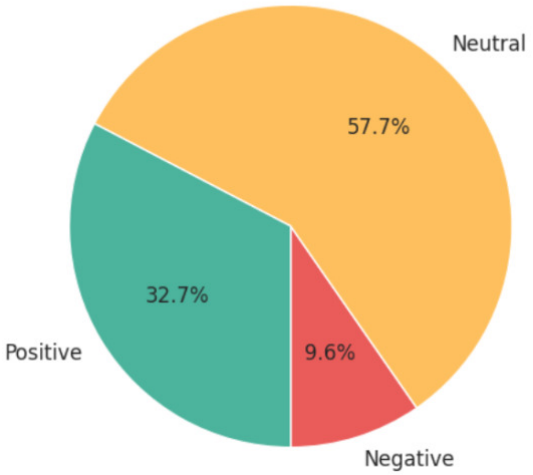

Figure 37 - Pie chart showing the distribution of Neutral, Positive, Negative sentiment

The pie chart presented in figure 37, shows that from the dataset of 15603 unique tweets collected between the 4th and 27th of February 2022, 57.7% (9000) of the tweets had Neutral sentiment, 32.7% (5098) had Positive and a percentage of 9.6% (1505) had Negative. This shows that there were more tweets with a Positive sentiment than a Negative Sentiment.

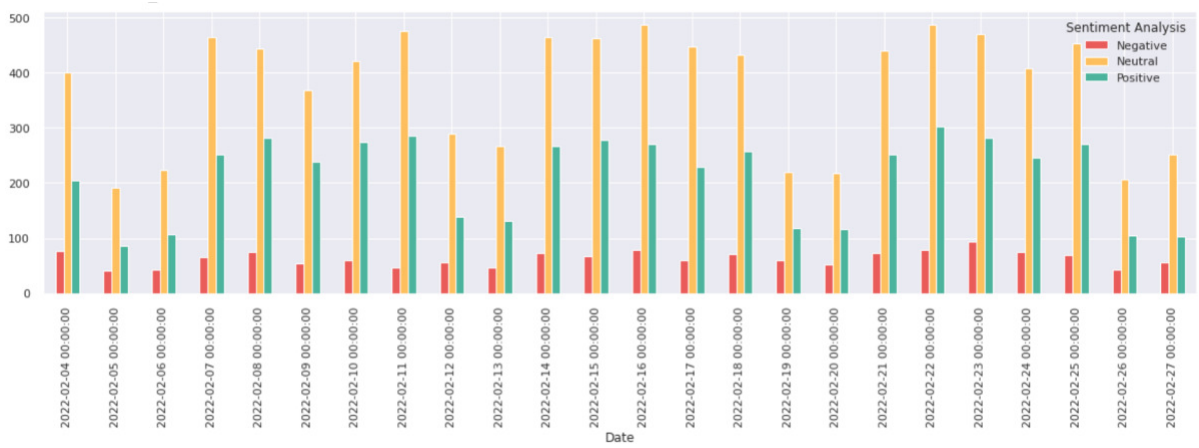

Figure 38 - Bar plot showing the number of Neutral, Positive and Negative tweets each day of collection

Figure 38, shows a bar plot of the Negative, Neutral and Positive sentiments over the collection period. As it can be seen, during all the days, the sentiments Neutral, Negative and Positive had the same analogy. The Neutral sentiment tweets were varying between approximately 200 and 500, the Positive tweets were varying between approximately 100 and 300 and lastly, the Negative ones were between approximately 40 and 100.

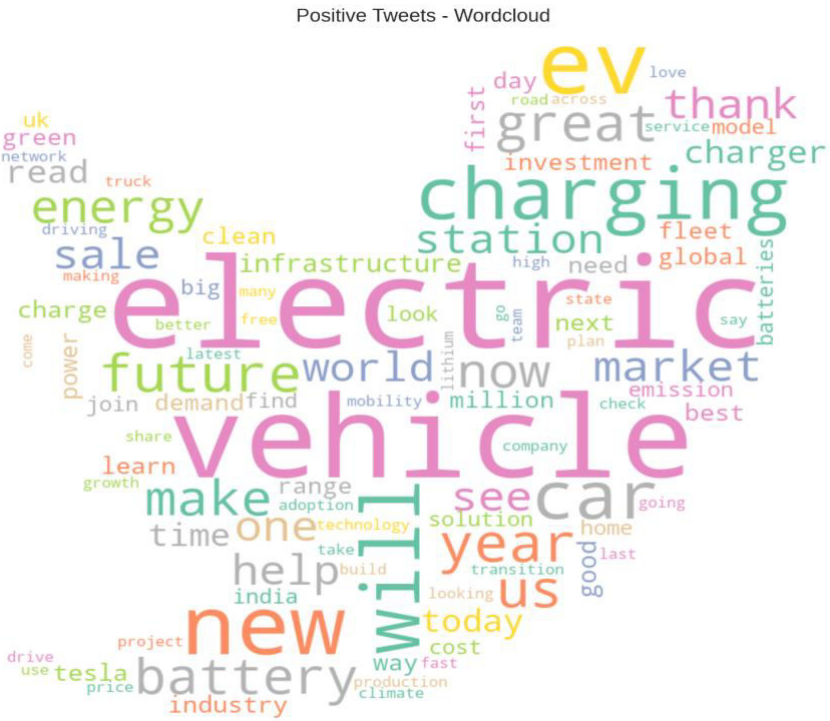

Figure 39 - Word cloud of the top 100 most common words in the tweets with Positive sentiment

Figure 39, shows a word cloud of the top 100 most frequent words found in the Positive tweets. Except for the words ‘electric’ and ‘vehicle’ that were unsurprisingly the most frequent words, some interesting words were ‘charging’, ‘station’, ‘battery’, ‘market’, ‘investment’, ‘sale’, ‘future’, ‘solution’, ‘transition’ and ‘infrastructure’. Additionally, words like ‘thank’, ‘great’ and ‘love’ demonstrate that there is a sign of positivity in these tweets.

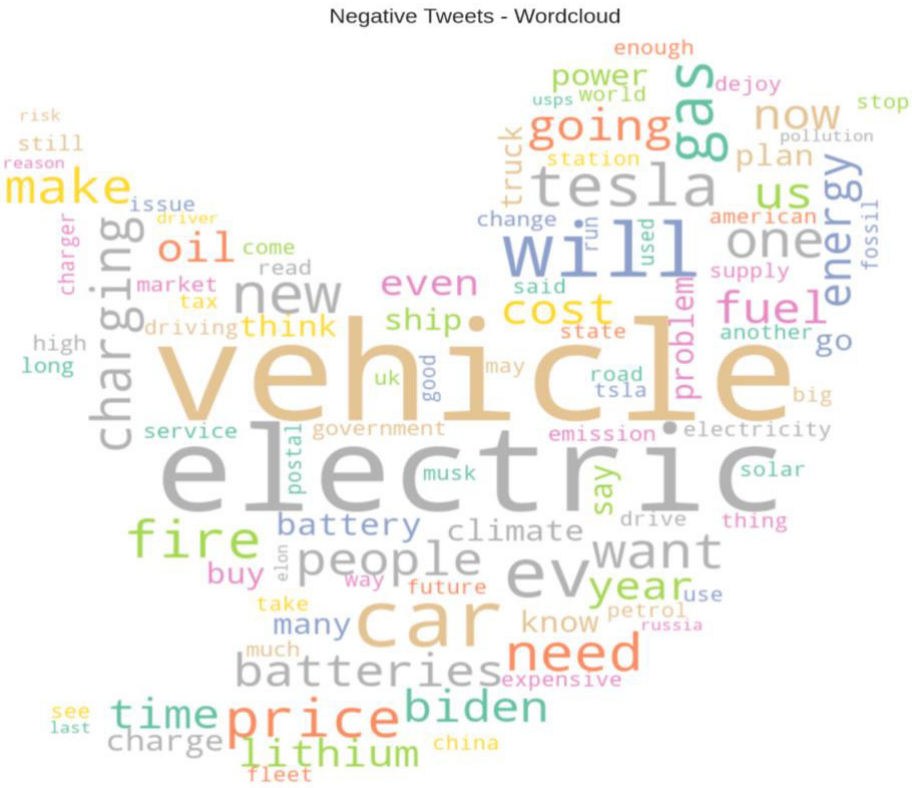

Figure 40 - Word cloud of the top 100 most common words in the tweets with Negative sentiment

Figure 40, shows a word cloud of the top 100 most frequent words found in the Negative tweets. Except from the words ‘electric’ and ‘vehicle’ that were predictably the most frequent words, some interesting words were ‘charging’, ‘fire’, ‘battery’, ‘tesla’, ‘cost’, ‘price’, ‘gas’, ‘biden’, ‘oil’ and ‘pollution’. Additionally, words like ‘risk’, ‘issue’ and ‘expensive’ demonstrate that there is a sign of negativity in these tweets.

The numbers of topics extracted from the positive topic model were 14 and each topic was contributed by the top 10 keywords with the highest probability distribution. In figure 41, the topics extracted from the positive topic model are shown and visualised as word clouds.

Figure 41 - Word clouds of the Topics extracted from the Positive Topic Model

As it can be seen in figure 41, not all topics could be interpreted solely from the topic keywords. Hence, in order to understand the topics better, I needed to check the tweets that represent each topic the most. Each document in an LDA model is composed of multiple topics, but only one is typically dominating. Therefore, I have found the top 5 tweets a given topic has contributed to the most and inferred the topic by reading the tweets. Some examples of the top tweets which represent each topic can be shown in Table 2.

Table 2 - Examples of one of the top 5 tweets which represent each topic extracted from the Positive Topic Model

Topic Number	Topic Percentage Contribution	Keywords	Text	Dates
0.0	0.2394	car, electric, read, model, top, motor, commercial, save, buy, fully	“Some great Superbowl #EV commercials. Which one was your favorite? 1. Kia's Robo Dog 2. GM's Dr. EV-il 3. Polestar's snarky ad 4. BMW with Zeus and Hera 5. Nissan with Eugene Levy 6. Chevy and the Sopranos”	2022-02-14

1.0	0.2494	charge, ev, station, charger, infrastructure, home, fast, road, network, state	“Yet another successful installation done by team EVRE! This time around, we installed our latest DC Fast Charger at @hiranandanigrp Panvel. This is a public charging station and can be accessed by any EV owner using our EVRE Mobile App. #EVRE #ElectricVehicles #India https://t.co/btM5j8u90k”	2022-02-16
2.0	0.2471	future, world, work, check, mobility, city, sustainable, transportation, partner, ready	“The team and I were honored to host & showcase our #EV vehicles to Hon. Environment Minister of GoM Sh. @AUThackeray ji. Thank you Aditya ji for your honest feedback, your words of encouragement and your commitment towards building a cleaner and greener tomorrow. @vikramsathaye https://t.co/2gVnzBI0Ep”	2022-02-19
3.0	0.2619	technology, industry, big, project, india, auto, development, set, ford, automotive	“PLI SCHEME PHASE 1, released 20 companies selected in this scheme phase 1 , all are SANDHAR TECHNOLOGIES LTD CLIENTS. BIG BOOM AHEAD. NEXT AS SANDHAR TECHNOLOGIES LTD IS AN OEM SUPPLIER, AND MOST ADVANCED ELECTRIC VEHICLES ANCILLARIES SUPPLIER , IT WILL BE IN TOP LIST IN PHASE https://t.co/Eq2FccIBEZ”	2022-02-12
4.0	0.2372	electric, vehicle, innovation, research, growth, half, cheaper, nissan, mercendes_benz, digital	“Blockchain Innovation => $EWT Energy Transition => $EWT Digitalization (DID) => $EWT Artificial Intelligence => $EWT Internet of things => $EWT Electric Vehicles (EV) => $EWT All roads lead to $EWT $EWT is literally a combination of all the best technologies in this decade! https://t.co/7bJQcPsEmm”	2022-02-10
5.0	0.2313	drive, range, love, late, article, full, experience, test, read, design	“Built for daily adventures and family getaways, the #Peugeot #eRifter's spacious boot capacity ensures nothing gets left behind…. Start your journey with the #eRifter at our #Kettering showroom, or click https://t.co/ggloREPA20 to learn more. #EV #PeugeoteRifter https://t.co/OkjRXOli1”	2022-02-09
6.0	0.1759	make, great, today, good, news, day, move, switch, thing, event	“Good news from @ZipCharge, taking some of the worry out of driving longer distances in #EVs. This might just be the news that reassures drivers enough that they're happy to make the switch to an #ElectricVehicle @CiTTimagazine https://t.co/CqXVEeCMBg”	2022-02-10
7.0	0.245	company, investment, million, include, plan, build, adoption, invest, create, forward	“Driven by @POTUS's economic strategy, major companies like Tritium, Intel, General Motors, and Boeing are investing $200B+ in U.S. manufacturing of semiconductors, electric vehicles and chargers, aircrafts, and batteries — creating good-paying jobs and lowering consumer costs https://t.co/askafvHHYA”	2022-02-09
8.0	0.2068	ev, time, share, find, uk, interest, show, sell, large, week	“EV sales data for December is in. More #EVs sold in December than any month to date. The number of EV models sold grew to 76, a 25% increase from the start of the year. Tesla took a huge 64% market share in December 2021. https://t.co/bcMEeJWHku”	2022-02-11
9.0	0.2475	ev, join, transition, consumer, back, step, revolution, climate, register, start	“We're supporting Arizona’s greater Phoenix region with the TE Activator. Join the next TE Tuesday webinar on March 1st to explore the social & economic implications of TE and how to drive an equitable transition. Register here: https://t.co/Y8Fxw2gzK0 #webinar #ElectricVehicles https://t.co/B19OikkzgN”	2022-02-18

10.0	0.2068	electric, vehicle, tesla, power, green, truck, gas, solar, bus, produce	“Welcome to a reimagined world of Born Electric vehicles. Electrifying presence & exhilarating performance brought to you by our team of global designers and experts. Coming soon July 2022 Follow Mahindra Born Electric to stay plugged-in. #bornelectricvision #ShivShaktiWahan https://t.co/EFOoNzPfSn”	2022-02-11
0.2681	0.2681	battery, year, market, sale, high, increase, demand, global, grow, price	“#China Electric Vehicle maker $JZXN has broken resistance and appears to be heading higher despite fears of #stockmarketcrash. This #ElectricVehicles sector is one that is expected to keep growing in order to keep up with demand. SRIDE $DRIV $TSLA $NIO $GOEV $LCID $LI $RMO $WKHS https://t.co/turfH9x4z1”	2022-02-14
12.0	0.1992	learn, fleet, solution, service, business, electrify, provide, launch, easy, free	“We are providing best #EV solutions for your existing or planned business with full software and hardware solution along with renowned EV #Charger Brands #electric #mobility #businessideas #sharktank #startups #mobbypark #mufc #evcharger #SupremeCourt #RocketBoys https://t.co/pJJrhXS7XY”	2022-02-05
13.0	0.2447	energy, clean, cost, emission, reduce, benefit, low, power, people, grid	“NSBA | White House Touts Clean Manufacturing - new actions across agencies to promote clean manufacturing and reduce emissions, including low-carbon production of steel and aluminum for electric vehicles, wind turbines, solar panels, and clean concrete. https://t.co/6gh6ptZX7b”	2022-02-18

After checking out the topics along with their tweets, with some domain knowledge and interpretation, I gave the topics a name. However, for some topics, it was difficult to accurately understand their meaning. Therefore, I have taken an approach that other people have taken, namely to discard the noisy topics and remain with the interpretable ones (Ylä-Anttila et al. 2021). Table 3 shows the names I have given to the topics and also the topics that I have discarded.

Table 3 - Labelled and Discarded Topics from Positive Topic Model

Topic Number	Label / Discarded
0	Commercials
1	Charging Stations
2	Greener Future
3	-Discarded-
4	-Discarded-
5	Electric Vehicle Journey
6	-Discarded-
7	Investments
8	Sales

9	Events
10	-Discarded-
11	Stock Market
12	Business Solutions
13	Sustainability

As it can be seen in Table 3, 4 out of 14 topics which were extracted have been discarded.

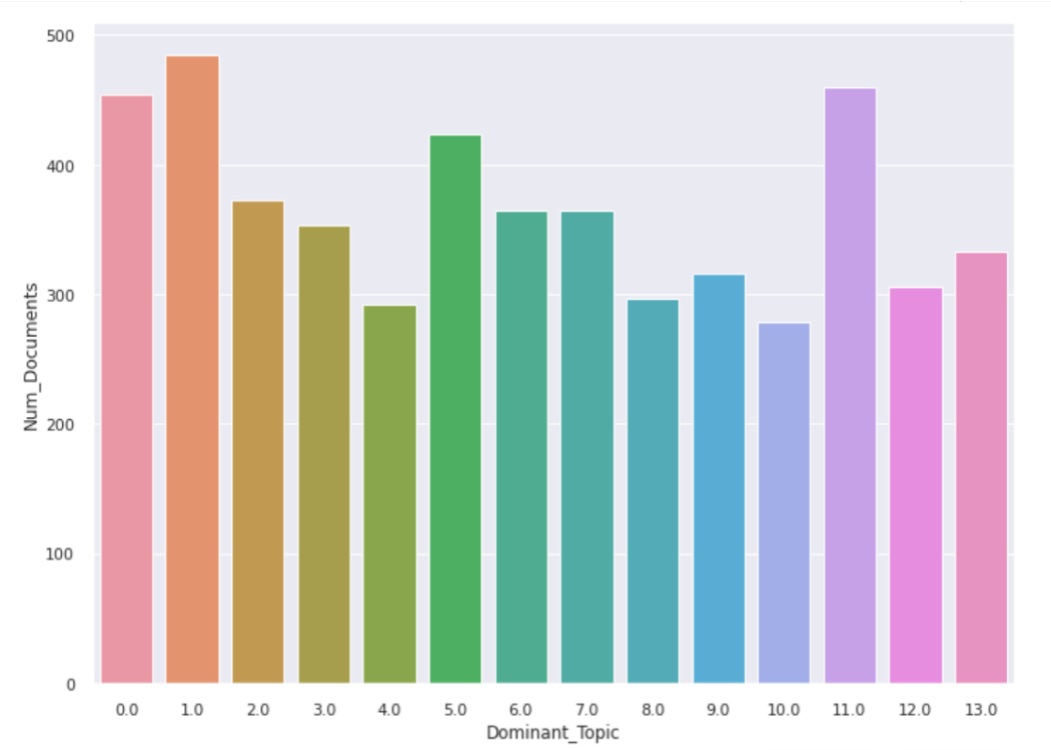

Figure 42 - Bar plot showing the number of tweets for every Positive dominant topic.

The bar plot in figure 42, shows the number of tweets at which a certain topic was dominant at. The topic that was dominant in most of the tweets was the 'Charging stations' (1.0) and the topic that was the least dominant was the 'Sales' (8.0), if we exclude topics 10.0 and 4.0 which have been discarded.

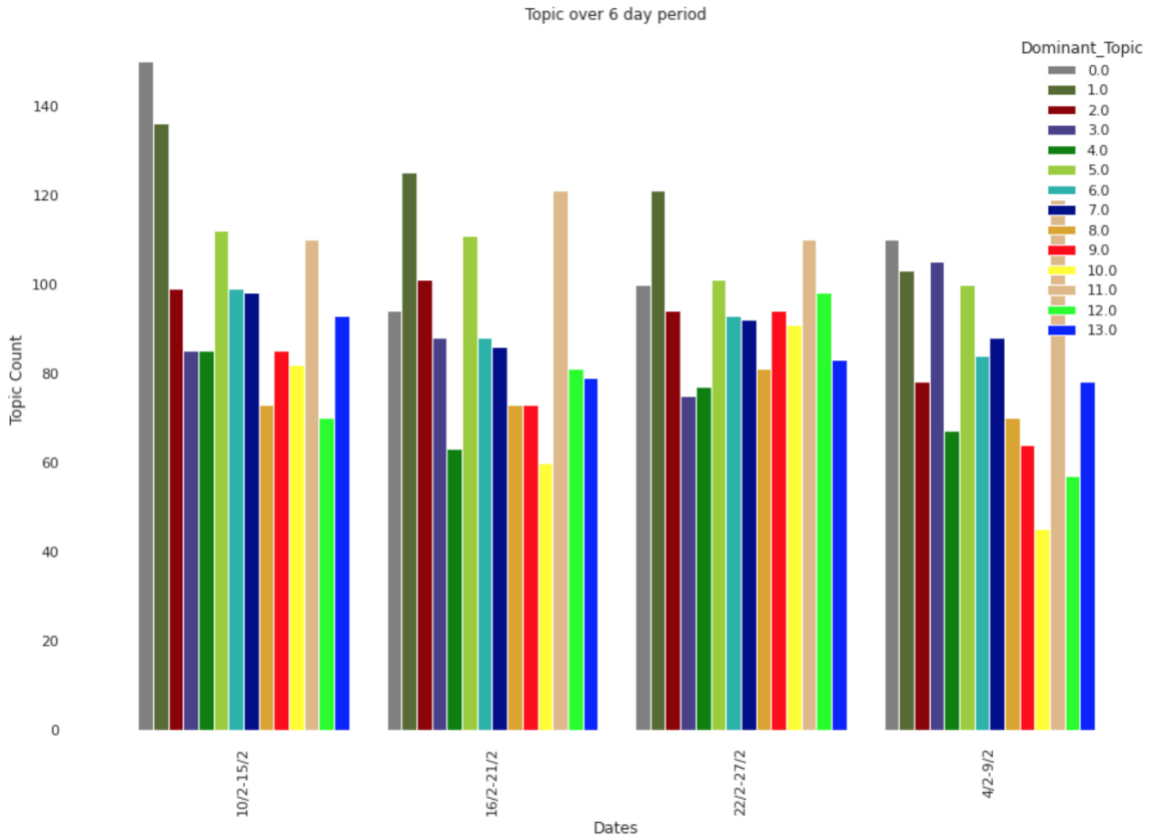

Figure 43 - Bar plot showing how topics extracted from the Positive Topic Model change over time (6 days periods)

Figure 43, shows a bar plot demonstrating how the topics from the positive topic model change over time. I have collected tweets for an overall of 24 days, so I split it into four periods of 6 days. During the 4th and 9th of February, the dominant topic appears to be the 'Stock Market' (11.0) and the least dominant the 'Business Solutions' (12.0), if we exclude topic 10.0 that has been discarded. During the 10th and 15th of February, the dominant topic appears to be the 'Commercials' (0.0) and the least dominant the 'Business Solutions' (12.0). During the 16th and 21st of February, the dominant topic appears to be the 'Charging Stations' (1.0) and the least dominant the 'Sales' (8.0) and 'Events' (9.0), if we exclude the topics 10.0 and 4.0 which have been discarded. During the 22nd and 27th of February, the dominant topic appears to be the 'Charging Stations' (1.0) and the least dominant the 'Sales' (3.0), if we exclude the topics 3.0 and 4.0 which have been discarded. An observation made was that all the topics were persistent and this possibly occurs due to the fact that the data were collected over a quite short period of time.

Some interesting topics which had a positive sentiment and seemed to be discussed regularly during the period of the data collection, are the topics of 'Sustainability', 'Greener Future' and 'Charging Stations'. According to research regarding greenhouse gas emissions from the manufacturing of lithium-ion batteries for electric vehicles, electric vehicles typically have much lower life-cycle greenhouse gas emissions than a typical car in Europe, even when assuming relatively high battery manufacturing emissions (Hall and Lutsey 2018). Therefore, this shows that electric vehicles can indeed lead to sustainability and a greener future and that can have the effect of people having a positive sentiment about electric vehicles.

The fact that the remaining topics such as 'Commercials', 'Sales', 'Events' and 'Business

Solutions', had positive sentiments is unsurprising, because usually their purpose is to positively affect people's views about a product and in this case, the electric vehicles.

The numbers of topics extracted from the negative topic model were 26 and each topic was contributed by the top 10 keywords with the highest probability distribution. In figure 44, the topics extracted from the negative topic model are shown and visualised as word clouds.

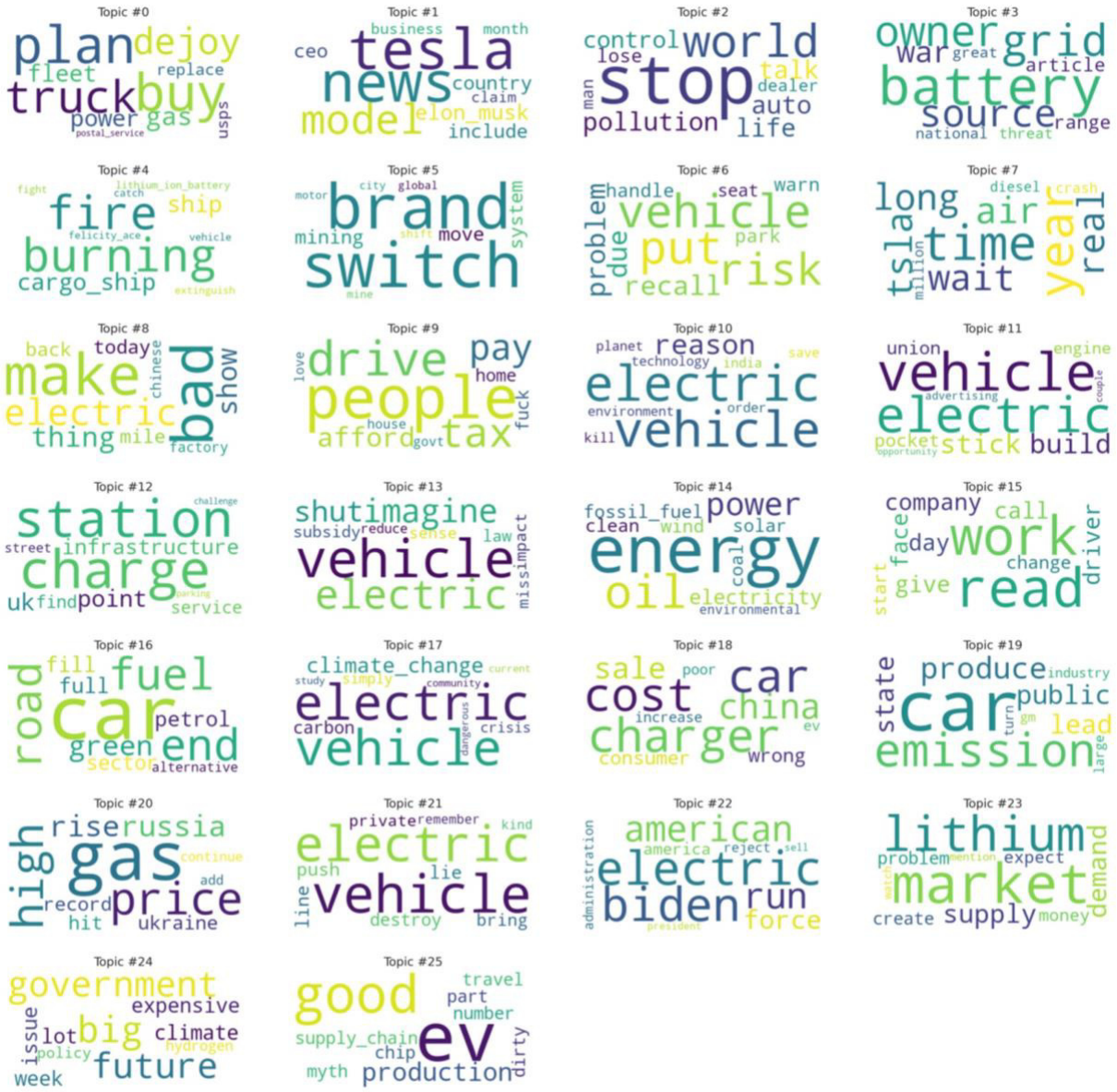

Figure 44 - Word clouds of the Topics extracted from the Negative Topic Model

As it has been observed from the positive topic model, not all topics could be interpreted merely from the topic keywords. This is also observed in the topics extracted from the negative topic model. Hence, I have followed the same approach as the one I have followed with the positive topic model, which was to find the top 5 tweets that a given topic has contributed to the most and infer the topic by reading that document. Some examples of the top tweets which represent each topic can be seen in table 4.

Table 4 - Examples of one of the top 5 tweets which represent each topic extracted from the Negative Topic Model

Topic Number	Topic Percentage Contribution	Keywords	Text	Dates
0.0	0.2117	buy, plan, truck, dejoy, gas, power, fleet, replace, usps, postal_service	“The White House and Environmental Protection Agency on Wednesday blasted Louis DeJoy and the US Postal Service for its decision to usps, postal service replace the vast majority of its aging vehicle fleet with gas-powered trucks, rather than electric vehicles.”	2022-02-04
1.0	0.2314	tesla, news, model, elon_musk, country, ceo,include, business, claim, month	“Business Insider - Tesla is under investigation after a South Korean regulator said it exaggerated mileage claims for vehicles including the Model 3: report https://t.co/Esp9moBISv https://t.co/8fy4h4ThoH”	2022-02-15
2.0	0.1427	stop, world, pollution, auto, talk, life, control, lose, dealer, man	“Wow: "Across Boston 1 in 5 cases of childhood asthma is estimated to be attributable to pollution, but in neighborhoods with the most traffic-related pollution...it can be up to 1 in 3." Similar playing out in other cities. So many reasons to switch to EVs https://t.co/zfUCfU9952”	2022-02-15
3.0	0.147	battery, grid, owner, source, war, range, article, national, threat, great	“The U.S. Has A Battery Problem That Could Turn Into A National Security Threat The transition to electric vehicles is going to require a lot of EV batteries, which the U.S. doesn't currently have. Our lack of batteries and the raw…..https://t.co/67bQCztOQ0 https://t.co/ibp7YwB2YE”	2022-02-24
4.0	0.2227	fire, burning, cargo_ ship, ship, lithium_ion_battery, fight, catch, extinguish, felicity_ace, vehicle	"Why so many container ships catch fire" "The cause: Lithium-ion batteries in electric vehicles have apparently caught fire - which complicated the extinguishing of the fire: "The ship is burning from one end to the other," said the port captain of Horta…."https://t.co/qoXAKU3Fqg	2022-02-19
5.0	0.1832	brand, switch, mining, move, system, global, shift, mine, city, motor	“The dirty side of Electric Vehicles. Nickel mining for the EV industry has contaminated drinking water close to Indonesia's largest nickel mine with unsafe levels of hexavalent chromium (Cr6), the cancer-causing chemical, made famous by Erin Brockovich. https://t.co/wbObCGVjT8“	2022-02-20
6.0	0.1827	vehicle, put, risk, recall, problem, due, park, handle. warn, seat	“Park Outside: Hyundai. Kia Recall Vehicles Due to Fire Risk Funny how vehicles filled with EXPLOSIVE GASOLINE/PETROL have never been a serious FIRE RISK, but 'save the environment' ELECTRIC VEHICLES can't seem to stop SELF COMBUSTING) Chevy BOLT &amp: KIAQ httos://t.co/HU6laLOJH1”	2022-02-09

7.0	0.1424	time, vear, lona, wait, air, tsla, real, diesel, million, crash	“Rimac Has Spent $20 Million Crashing A Bunch Of Never Hypercars. The final crash test has been completed, so US customers won't have long to wait now. #crash #electricvehicles #supercars #video Read: https://t.co/gU59ZCvLQd https://t.co/QHIE89qM2H”	2022-02-18
8.0	0.1243	make, bad, electric, thing, show, today, back, mile, factory, chinese	“EVs are bad for the environment and I can prove it. EVs are made by smelting Lithium in China, which makes CO2. The factories are powered with COAL. To make EVs, the create the CO2 in one day as driving a F-150 for 100,000 miles. That's bad. #EVs”	2022-02-22
9.0	0.1713	people, drive, tax, pay, afford, home, fuck, house, govt, love	“What a surprise, the treasury says that motorists need taxing more, why? because more people are driving electric vehicles,more like people can't afford to drive especially when Politicians treat taxpayers money like their own and keep giving it away to people who don't need it.“	2022-02-04
10.0	0.1579	electric, vehicle, reason, environment, save, kill, planet, technology, india, order	“However, it's in Europe's best interest to be looking like they care about Ukraine because of its untapped resources the global BULLY squad, needs to sustain its electric vehicles etc. Kill the Ego to save the soil of the earth. Kill the Ego save the soul of the man https://t.co/pLSxrrAhVt”	2022-02-23
11.0	0.10.47	vehicle, electric, build, stick, pocket, union, engine, advertising, opportunity, couple	“Why are lawmakers allowing monopoly of an #American company& industry? Whose lining whose pockets to keep these archaic decisions in play? #America deserves better& needs to stop letting CORPORATE politics get in the way of its own citizens. #Tesla #Automotive #ElectricVehicles https://t.co/hvfuzlV6h”	2022-02-21
12.0	0.1629	charge, station, point, uk, infrastructure, service, find, street, challenge, parking	“#Birmingham residents who don't have off-street parking will find it more challenging to charge without a hub nearby. Complete our survey to explain the challenges you face to make the switch to #EV https://t.co/QhtweV41z https://t.co/G90Jc65JeM”	2022-02-07
13.0	0.11	vehicle, electric, imagine, shut, subsidy, law, impact, sense, miss, reduce	“I sure hope those are all electric vehicles. Electric tanks. Also, the soldiers inside need to be masked, vaxxed &boosted plus practicing social distancing. Oh my GOD are those soldiers carrying AR15s? We should just pass a law against this. Putin would respect the law game over https://t.co/UyhiK3ChLu“	2022-02-25
14.0	0.1841	energy, oil, power, electricity, fossil_fuel, solar, clean, wind, coal, environmental	“This is a false choice between dependence on Putin/Saudi energy/oil and making our own. We can do better: -Electric vehicles -Solar -Wind It is time to phase out coal, nuclear, and fracking and do what we can to save the planet. https://t.co/pFE8c4VUFz”	2022-02-26
15.0	0.1603	work, read, company, driver, give, face, call, day, change, start	“The tool & die shop I work for had to turn away $1.2m in new work because we're already buried in over $6m in new	2022-02-19

			work. All new projects are for electric vehicles. Not enough die shops or tool makers anymore. Won't be long and they'll have to give CNC and tool makers blank checks“	
16.0	0.1405	car, fuel, road, end, green, petrol, full, fill, sector, alternative	“Went to the petrol station to jet wash the cemetary mud off my car. Not been to a petrol station in about a year now (I drive an #EV). Complete madness - the cost of fuel, the lining up to fill and then to pay and the act of filling your car with an inflammable liquid!”	2022-02-27
17.0	0.1116	electric, vehicle, climate_change, carbon, simply, crisis, community, study, current, dangerous	“Sixteen Fisker Karma electric vehicles caught fire & burned to the ground after being submerged by saltwater from Hurricane Sandy's storm surge. The vehicles were submerged when storm surge beached the port flooding the luxury electric vehicles and other cars parked in the port. https://t.co/5RVy1n07Ee”	2022-02-16
18.0	0.1971	cost, charger, car, china, sale, wrong, consumer, increase, poor, ev	If #EVs aren't accessible to EVeryone, then we're doing it wrong. "..public charger access is lower in block groups with below-median household incomes and in those with a Black & Hispanic majority populations." https://t.co/WVbvL9RmAS	2022-02-15
19.0	0.2392	car, emission, produce, public, lead, state, industry, gm, large, turn	“Elon Musk in email to #CNBC: "Biden has pointedly ignored #Tesla at every turn and falsely stated to the public that GM leads the electric car industry, when in fact Tesla produced over ---> 300,000 electric vehicles last quarter and #GM produced 26." $GM $TSLA #ElectricVehicles https://t.co/79Mr6HnRLu”	2022-02-23
20.0	0.1302	gas, price, high, russia, rise, ukraine, record, hit, continue, add	“Can anyone say Electric vehicles? How Russia's invasion of Ukraine will send already high gas prices higher https://t.co/8khxABL3ke via @Yahoo”	2022-02-27
21.0	0.092	vehicle, electric, push, lie, line, destroy, bring, private, remember, kind	“A setback in Rio Tinto's #lithium development project in Serbia has exposed the overall fragility of Europe's outlook for bringing #BatteryRawMaterial material sourcing closer to home. Read the full story: https://t.co/Hfk7WJb2yN #FastmarketsEnergyTransition #EV https://t.co/krBL2j6ljl”	2022-02-07
22.0	0.1453	electric, biden, american, run, force, america, reject, administration, president, sell	“Joe Biden went to war against the American energy industry on day one. Joe Biden took several steps to ensure high gas prices because Electric vehicles look stupid when gas prices are low and Biden wants to force everyone into electric cars. Biden forced high energy cost. https://t.co/7Y94tXQGn4”	2022-02-26
23.0	0.1299	market, lithium, supply, demand, problem,	“The market potential for lithium is forecasting extreme growth! The supply of lithium has not and will continue to not be	2022-02-09

		create, expect, money, mention, watch	able to meet the demands. #lithium #mining #lithium battery #ev #electricvehicle https://t.co/jqbrKAjIAN”	
24.0	0.1574	big, government, future, expensive, issue, climate, lot, week, policy, hydrogen	“The biggest issue with what is happening is this line - "But the Department for Transport's key target is to increase the uptake of electric vehicles, so our report is providing a disincentive to that." Focus should not be on a single solution when there are many others. https://t.co/482e79b4MQ”	2022-02-18
25.0	0.1572	ev, good, production, supply_chain, part, number, myth, dirty, travel, chip	“EV production emissions myths found to be just that: "The supply chain for combustion vehicles is just so dirty that electric vehicles can't surpass 2022-02-24 them, even when you factor in indirect emissions." https://t.co/ZOAI03gnsf“	2022-02-24

After checking out the topics along with their tweets, I have taken the same approach as I had done so with the tweets that had a positive sentiment. I have labelled the interpretable topics with the best suitable name and I have discarded the topics that were not interpretable. Table 5 shows the names I have given to the topics and also the topics I have discarded.

Table 5 - Labelled and Discarded Topics from Negative Topic Model

Topic Number	**Label / Discarded**
0.0	Replacement of the US Postal Service ageing vehicles to gas-power trucks
1.0	Tesla under Investigation in South Korea
2.0	Pollution
3.0	Issues related to EV batteries
4.0	Fire on Cargo ship with electric vehicles
5.0	Mining Challenges
6.0	Fire risk of electric vehicles when parked
7.0	-Discarded-
8.0	-Discarded-
9.0	Government Tax
10.0	-Discarded-
11.0	-Discarded-
12.0	Charging stations challenges
13.0	-Discarded-

14.0	Switching to renewable energy
15.0	-Discarded-
16.0	-Discarded-
17.0	-Discarded-
18.0	Charger Access
19.0	-Discarded-
20.0	Gas prices increase due to Russia-Ukraine conflict
21.0	-Discarded-
22.0	Joe Biden wants to force electric vehicles
23.0	Supply of Lithium does not meet the demands
24.0	-Discarded-
25.0	Myths relating to EVs

As it can be seen in table 5, 11 out of 26 topics which were extracted have been discarded.

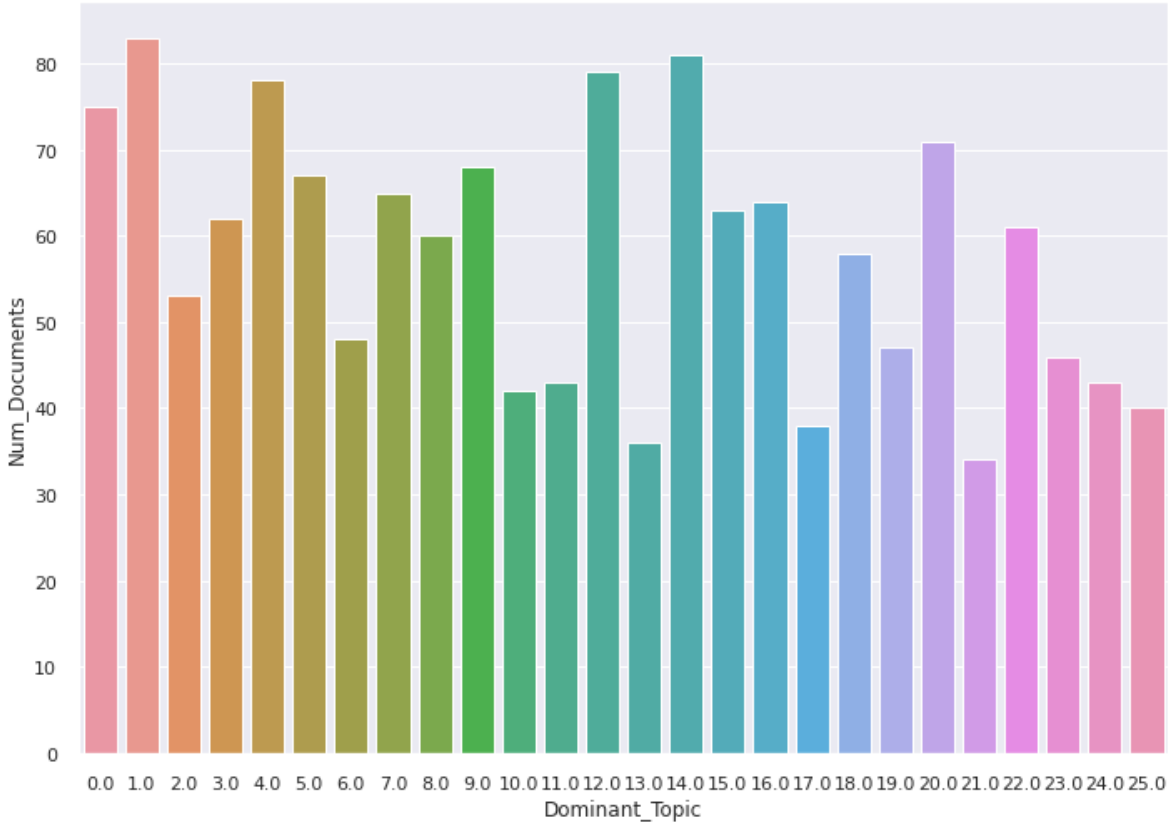

Figure 45 - Bar plot showing the number of tweets for every Negative dominant topic.

The bar plot in figure 45, shows the number of tweets in which a certain topic was dominant at. The topic that was dominant in most of the tweets was the 'Tesla under investigation in South Korea' (1.0) and the topic that was the least dominant was the 'Myths relating to EVs' (25.0), if we exclude the topics 21.0, 13.0, 17.0 which have been discarded.

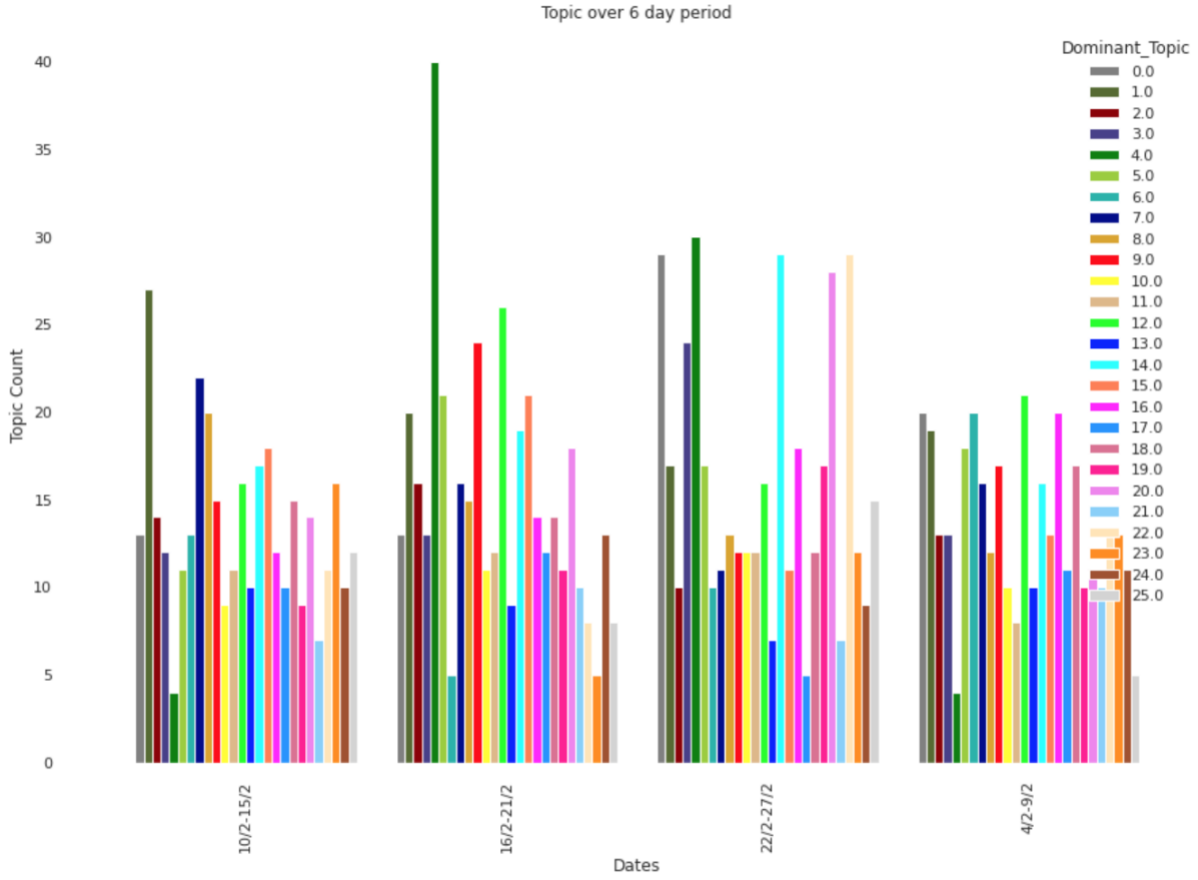

Figure 46 - Bar plot showing how topics extracted from the Negative Topic Model change over time (6 day periods)

Figure 46, shows a bar plot demonstrating how the topics from the negative topic model change over time. The days were split again into four periods of six days. During the 4th and 9th of February the dominant topic appears to be the 'Charging station challenges' (12.0) and the least dominant topic the 'Fire on Cargo ship with electric vehicles' (4.0). During the 10th and 15th of February the dominant topic appears to be the 'Tesla under Investigation in South Korea' (1.0) and the least dominant, once again the 'Fire on Cargo ship with electric vehicles' (4.0). During the 16th and 21st of February, the dominant topic appears to be the 'Fire on Cargo ship with electric vehicles' (4.0) and the least dominant topic the 'Fire risk of electric vehicles when parked' (6.0) and 'Supply of Lithium does not meet the demands' (23.0). During the 22nd and 27th the dominant topic appears to be again the 'Fire on Cargo ship with electric vehicles' (4.0) and the least dominant topic the 'Pollution' (2.0) and 'Fire risk of electric vehicles when parked' (6.0), if we exclude the topics 17.0, 13.0, 21.0, 24.0 which have been discarded. Moreover, it can be seen that the topics 'Replacement of the US Postal Service ageing vehicles to gas-power trucks' (0.0), 'Switching to renewable energy' (14.0), 'Gas prices increase due to Russia-Ukraine conflict' (20.0) and 'Joe Biden wants to force electric vehicles' (22.0) have a high increment compared to the previous periods. In addition, most of the topics seem to be persistent over time. As mentioned for the topics extracted from the positive topic model, this is because the data were collected in a quite short period of time. One topic that was not very

discussed at the beginning and then it was converted to one of the most discussed topics, is the 'Fire on Cargo ship with electric vehicles' (4.0).

An interesting topic that seems to be widely discussed during the period between 16th and 27th of February, is the 'Fire on Cargo ship with electric vehicles'. According to Sakellariou (2022), on February 18th a ship carrying around 4,000 electric and non-electric vehicle models from brands such as Audi, Porsche, Bentley, Lamborghini, and Volkswagen, caught fire near the coast of the Azores, Portugal. The exact cause of the incident has not been determined, but it is believed that a lithium battery commonly used in electric cars may have caught fire. This incident might have caused people to have a negative sentiment about electric vehicles. Another noteworthy topic, is the topic of 'Mining Challenges'. According to an investigation by Guardian regarding nickel mining and the electric vehicle sector, a source of drinking water near one of Indonesia's largest nickel mines is contaminated with dangerous amounts of hexavalent chromium (Firdaus and Levitt 2022). The article was published on the 19th of February and as it can be seen in Figure 46, there is an increase of tweets that had 'Mining Challenges' as a dominant topic during the period 16th and 21st of February. Therefore, this investigation might have had a negative impact on people's sentiments. Furthermore, a topic which was expected to be found was the 'Charging station challenges', which has also been the most dominant topic between the 4th and 9th of February. One of the main reasons people are negative about electric vehicles, is not the vehicles themselves but the recharging. They are afraid that the vehicles do not have sufficient energy storage to cover the planned distance to the intended destination. This phenomenon is called "range anxiety" and can be eliminated by building an appropriate and efficient EV charging infrastructure (Tridens Technology 2022). Lastly, the topic 'Gas prices increase due to Russia-Ukraine conflict' impacted people's sentiment but it did not directly affect people's sentiment about electric vehicles. According to Energy Live News, energy prices have risen as a result of the ongoing crisis in Ukraine, as sanctions are imposed on Russian oil and gas companies. Hence, electric cars seemed to be the next industry to suffer. Without the Russian market, there is not enough infrastructure for building Nickel batteries and this will require a heavier focus on Lithium batteries which are less efficient (Bose 2022). Also, this situation interconnects with the topics 'Mining Challenges' and 'Pollution', because the mining of Lithium causes pollution to the environment.

4.1 Summary of Results

After analysing the graphs and findings presented above, it can be seen that during the collection period the data mostly had a neutral sentiment towards the electric vehicles. However, it is important to note that the percentage of the tweets with a positive sentiment was much higher than the percentage of the negative ones.

Even before applying the topic models to the positive and negative tweets, some frequent words in the tweets were an indication of what was being discussed. Some of these words in the positive tweets were about charging stations, batteries, investments, sales, market, transition etc. and some words in the negative tweets were about charging, batteries, cost, fire, price, gas, pollution, Biden etc.

After applying the topic models to the positive and negative tweets respectively, some interesting topics were extracted. Regarding the tweets which had a positive sentiment, some of the critical topics that were identified included Sustainability, Greener Future, Charging Stations, Investments, Sales, Stock Market and Business Solutions. These topics could have

impacted people to have a positive sentiment about electric vehicles. On the other hand, regarding the tweets which had a negative sentiment, some of the topics that were extracted included charging stations challenges, issues related to EV batteries, pollution, mining challenges, switching to renewable energy, charger access, the risk of fire of electric vehicles when parked, myths relating to EVs and the lack of lithium supply. What is interesting is that some other topics extracted from the model, were real-time events that took place during the data collection. Such topics included the replacement of the US Postal Service ageing vehicles to gas-power trucks, the investigation of the Tesla in South Korea, an incident where a cargo ship with electric vehicles caught fire, the increment of gas prices due to the conflict between Russia and Ukraine and Joe Biden's promotion of electric vehicles. These topics and events could lead people to directly or indirectly have a negative sentiment about electric vehicles.

4.2 Evaluation of Results

4.2.1 Evaluation of Collecting Tweets

Using the python library Tweepy for the connection to the Twitter API seemed to be a very good choice. It operated successfully in terms of filtering and retrieving the tweets related to electric vehicles and the tweets could easily be saved into a readable Excel file. Nevertheless, I have not found the documentation very helpful, as there were no examples of how the functions of Tweepy could be used and as a result, I had to search in other resources. However, I have had an 'Elevated' access to Twitter's API, and I could not make requests of more than 100 tweets before the last 7 days which was not very efficient. In addition, only a small number of the retrieved tweets had a location attached and as a result, an analysis of the tweets based on locations was not possible.

4.2.2 Evaluation of Sentiment Analysis Tools

The sentiment analysis tool that I used was a roBERTa-base model developed by Cardiff NLP research group. According to the results, the model seems to be achieving the task of providing the sentiment of the tweets. It has successfully returned 'Neutral' when the tweets lack sentiment, 'Positive' when the tweets had positive sentiment and 'Negative' when the tweets had negative sentiment. This was checked manually by looking at the tweets and it is also distinct from the topics extracted from the group of tweets that had positive sentiment and the group of tweets that had negative sentiment. For example, a topic that was extracted from the topic model which was applied to the positive tweets and was clearly positive is the 'Sustainability' and a topic that was extracted from the topic model which was applied to the negative tweets and was clearly negative is the 'Charging stations challenges'. Moreover, I have tried using the lexicon-based sentiment analysis model called 'Text-blob' and the results had a huge difference. That model was classifying most of the tweets as positive instead of neutral, because it was just checking the words without having an understanding of the tweets. All in all, I believe that the model I used was a very good choice for the task of sentiment analysis.

4.2.3 Evaluation of Topic Modeling Tools

The topic model I used was the LDA Mallet model. There is a number of practices which people commonly use in order to evaluate topic models, including human judgment, quantitative metrics – coherence calculations and the mix of these two approaches

(Rabindranath 2022). I chose the practice which combines the two approaches, as the coherence score was already calculated when I chose the models of negative and positive tweets with the best possible combination of coherence score and number of topics. For the tweets which had a positive sentiment, I chose the topic model with a number of topics K = 14 with a coherence score of 0.3658 which is relatively low (Pelgrim 2021). However, 10 out of 14 topics extracted from the model were interpretable. Nevertheless, not all the interpretable topics gave an indication of whether people's sentiments were impacted by them. For the tweets which had negative sentiment, I chose the topic model with a number of topics K = 26 with a coherence score of 0.5089 which is okay (Pelgrim 2021). Though, 15 out of 26 topics extracted were interpretable. Some of the interpretable topics discovered real-time events, for example, the incident of the cargo ship, the conflict between Russia and Ukraine etc. which directly or indirectly affected people's sentiment.

The topic model, even though it extracts a set of keywords as topics, cannot identify what exactly the topics are. Therefore, I needed to check the topics/set of keywords manually one by one in order to label them or discard them if they did not make sense. I have run the models many times with a different number of topics to get the most interpretable topics. Another issue with the LDA topic modeling, is that sometimes you may know that some of the tweets discuss a topic which you can identify, but the model cannot. There is no way to tell the model that certain words belong together.

All in all, the model extracted some very interesting topics which were related to electric vehicles and had an impact on people's sentiment during the period of the data collection, and therefore I believe that it achieved its objective.

4.2.4 Evaluation of Analysis of Results

Firstly, the objective relating to the analysis of the tweets' sentiment was accomplished. The percentages of the tweets which had Positive, Negative and Neutral sentiments were successfully identified and also an analysis of how the sentiments changed overtime was completed. In terms of analysing the extracted topics which were being discussed during the data collection, research about these topics was undertaken at that time, in order to see if there were events that impacted peoples' sentiment. However, line graphs instead of bar plots could have presented more clearly which topics were being discussed over time and more accurate insights would have been extracted. Also, the analysis of sentiment and the topics based on different locations around the world were not feasible, as not all the tweets had a location attached.

5. Conclusions

The aim of this project was to analyse the sentiment that people have and the topics that people are talking about on Twitter. To achieve this aim, a set of objectives had to be accomplished which included:

1. Retrieving the data from Twitter and creating a dataset.
2. Defining the sentiment of the tweets.
3. Pre-processing the tweets.
4. Identifying and interpreting the topics that are being discussed.
5. Analysing and visualising the findings to gain insights.

The first aim of retrieving data from Twitter and creating a tweet dataset was achieved. The tweets were collected by using the Twitter API and the python library Tweepy. The dataset was created in an Excel file, which included the tweets' id, the content of the tweets, the date and time and the location if there was any. I have collected 15603 unique tweets in total between the 4th and 27th of February 2022. I believe that this was a sufficient number of tweets which enabled me to analyse people's sentiments and the topics that are being discussed about electric vehicles, because the results of the models were interpretable. However, due to the fact that I had 'Elevated' access to the Twitter API, the method I used to collect the data included the manual setting of the date and time and retrieval of 100 tweets per request, something that was not very efficient. Also, not every day of the collection had the same number of tweets because there were duplicates in the tweets that had to be removed.

The second aim of classifying the tweets' sentiment was also achieved. The cleaning of the data was done successfully and the sentiment analysis model was applied to the cleaned tweets. The results of the sentiment analysis model seemed to be very reasonable as I had checked manually some of the tweets of each sentiment. Additionally, the topics which were extracted from both positive and negative topic models seemed to be a good indication that the sentiment analysis model worked, due to the fact that their classification as positive and negative was as reasonably expected.

The third aim of pre-processing the tweets and preparing them for topic modeling was successfully achieved as well. The tweets were tokenized, the stopwords were removed and bigrams and trigrams were created in the text. Moreover, dictionaries and bags of words (BOW) were successfully created to be passed to the LDA Mallet model.

The fourth aim of identifying and interpreting the topics that were being discussed during the time of the data collection was achieved. The topic models successfully extracted topics from the positive and negative tweets respectively. However, not all the topics were interpretable and some of them had to be discarded. Some very interesting real-time events were identified, which could probably have had a direct or indirect impact on people's sentiment about electric vehicles, including the incident of the cargo ship and the conflict between Russia and Ukraine.

The final aim of analysing and visualising the findings of the models was also achieved. By using tools and libraries in Python, I was able to analyse and visualise the raw data but also the results of the models. In terms of the raw data, bar plots were created showing the number of tweets collected daily, the 20 most used hashtags and the 20 most used mentions and a word cloud showing the frequent words. In terms of sentiment analysis of the tweets, a pie chart was created showing the percentages of each sentiment, a bar plot showing how sentiment changed day by day and word clouds showing the top frequent words in both Positive and Negative tweets respectively. Lastly, for the topics extracted from the models, word clouds were used to visualise the set of keywords for each topic, bar plots to visualise how topics changed over time and how many tweets each topic was dominant at.

Overall, the results showed that the sentiment of people regarding electric vehicles on Twitter was mostly neutral. However, the tweets with positive sentiments were more than the negative ones. The topics extracted from the topic models provided an indication of what was being discussed regarding electric vehicles during the period of collection, and why people had positive and negative sentiments. The results of this project do not coincide with the results found in the other studies reviewed. That is mainly because there is not much literature at the moment on sentiment analysis and topic modeling in the discussion about electric vehicles on

Twitter, especially during the time of my data collection. The only result that could be said to match with the study of Suresha and Kumar Tiwari (2021), which was the only study about electric vehicles reviewed, is that 'Tesla' was one of the top hashtags Twitter users tweeted while sharing tweets related to electric vehicles. This project has been able to successfully accomplish an analysis on the sentiment of the people and the topics that are being discussed on Twitter regarding electric vehicles. On that account, the findings of this project could be used by policy makers or the government to make decisions concerning campaigns which are run to inform the public about EVs.

6. Future work

Within the given timeframe, I believe I was able to conduct a good analysis of the discussion related to electric vehicles on Twitter. However, if I had been given more time, there are some things I would have improved. Firstly, an improvement I would make, even though I have attempted to do so, is to get an 'Academic Research' access which can give access to Twitter's full archive search, that provides Tweets from as early as 2006 and can deliver up to 500 Tweets per request. In this way, I would have retrieved much more tweets without putting effort into changing the date and time repeatedly. Moreover, it would have been more beneficial to collect a larger corpus of tweets, over a longer period of time than just 24 days, and experiment with a more diverse dataset, as this could have improved the accuracy of the findings and therefore, a better analysis would have been conducted. Also, I believe that the retrieval of an equal number of tweets every day would have contributed to more accurate results. Secondly, I would get more tweets that are geo-tagged, in order to perform an analysis based on locations. It would have been interesting to explore the sentiment that people have and what topics related to electric vehicles are being discussed in different countries. Another improvement I would make is to retrieve the id of the users, so that they could be later used in the analysis for the purposes of observing if the same users are posting. In addition, if more time was given, I would have tried various sentiment analysis models and different topic models in order to compare and find the ones that give the best results and are more suitable for the project.

In the near future, when the transition to electric vehicles progresses even further, it would be interesting to see how the sentiment of people and the topics that are being discussed regarding electric vehicles will change.

7. Reflections on learning

When I decided to choose this project, I was aware that it would be a challenge. This is because I did not have any knowledge regarding the field of data science and inevitably, a lot of time researching and learning would be required. I had never worked with Natural Language Processing techniques such as sentiment analysis and topic modeling before doing this project and therefore, I was very curious and excited to have an opportunity to learn about them and enhance my knowledge on the field. In addition, I had never used an API in the past and working with Twitter API in this project gave me the chance to learn more about how they work and how to use them. On top of learning these new skills, through the project I was able to enhance my skills in data manipulation, processing and analysis using Python. Particularly, I have learned to work with the Pandas and use NLP libraries such as the Natural Language Toolkit and Gensim, which are extremely useful for NLP tasks and techniques.

Having weekly meetings with my supervisor allowed me to receive continuous and useful feedback on the project's progress and stay on track. Organising a large-scale project like this, helped me to significantly enhance my planning and time management skills. Overall, I believe that I have managed to handle my time very well, as I was able to finish this project on time, while simultaneously preparing another coursework for the purposes of a different module.

Finally, through researching and conducting literature reviews, I have enhanced my research skills and I am more confident to perform such a research on future projects.

8. References

AltexSoft 2018. Sentiment Analysis: Types, Tools, and Use Cases. Available at: https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/ [Accessed: 22 March 2022].

Beri, A. 2020. Stemming vs Lemmatization. Available at: https://towardsdatascience.com/stemming-vs-lemmatization-2daddabcb221 [Accessed: 20 April 2022].

Bitext 2017. Lemmatization to enhance Topic Modeling results. Available at: https://blog.bitext.com/lemmatization-to-enhance-topic-modeling-results#:~:text=As%20we%20can%20see%20in,a%20better%20understanding%20of%20it [Accessed: 19 April 2022].

Boon-Itt, S. and Skunkan, Y. 2020. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. JMIR Public Health and Surveillance 6(4), p. e21978. doi: 10.2196/21978.

Bose, K. 2022. *Why is Russia's invasion of Ukraine halting EV production?*. Available at: https://www.energylivenews.com/2022/03/14/why-is-russias-invasion-of-ukraine-halting-ev-production [Accessed: 29 April 2022].

Canalys 2022. Global electric vehicle sales up 109% in 2021, with half in Mainland China. Available at: https://www.canalys.com/newsroom/global-electric-vehicle-market-2021 [Accessed: 21 March 2022].

Castelvecchi, D. 2021. Electric cars and batteries: how will the world produce enough?. Available at: https://www.nature.com/articles/d41586-021-02222-1 [Accessed: 29 March 2022].

Climate Nexus [no date]. Impacts from Mining Lithium, Cobalt, and Other Materials for Electric Car Batteries | Climate Nexus. Available at: https://climatenexus.org/climate-issues/energy/electric-car-batteries-impacts [Accessed: 29 March 2022].

Firdaus, F. and Levitt, T. 2022. *'We are afraid': Erin Brockovich pollutant linked to global* electric car boom. Available at: https://www.theguardian.com/global-development/2022/feb/19/we-are-afraid-erin-brockovich-pollutant-linked-to-global-electric-car-boom [Accessed: 26 April 2022].

Hidayatullah, A. et al. 2018. Twitter Topic Modeling on Football News. 2018 3rd International Conference on Computer and Communication Systems (ICCCS). doi: 10.1109/ccoms.2018.8463231.

Hugging Face [no date]. cardiffnlp/twitter-roberta-base-sentiment · Hugging Face. Available at: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment [Accessed: 23 March 2022].

Hugging Face [no date]. roberta-base · Hugging Face. Available at: https://huggingface.co/roberta-

base#:~:text=Model%20description,in%20a%20self%2Dsupervised%20fashion [Accessed: 23 March 2022].

Hugging Face [no date]. Transformers, what can they do? - Hugging Face Course. Available at: https://huggingface.co/course/chapter1/3?fw=pt [Accessed: 4 April 2022].

Hugging Face [no date]. Transformers. Available at: https://huggingface.co/docs/transformers/index [Accessed: 23 March 2022].

Jelodar, H. et al. 2018. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications 78(11), pp. 15169-15211. doi: 10.1007/s11042-018-6894-4.

Johnson, D. 2022. POS Tagging with NLTK and Chunking in NLP [EXAMPLES]. Available at: https://www.guru99.com/pos-tagging-chunking-nltk.html [Accessed: 20 April 2022]

Kapadia, S. 2019. Evaluate Topic Models: Latent Dirichlet Allocation (LDA). Available at: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 [Accessed: 30 April 2022].

Kemp, S. 2022. Digital 2022: Global Overview Report — DataReportal – Global Digital Insights. Available at: https://datareportal.com/reports/digital-2022-global-overview-report [Accessed: 9 May 2022].

Kosaka, M. 2020. Topic Modeling and Sentiment Analysis on Amazon Alexa Reviews. Available at: https://towardsdatascience.com/topic-modeling-and-sentiment-analysis-on-amazon-alexa-reviews-81e5017294b1 [Accessed: 30 March 2022].

Mal, M. 2021. Lemmatization Vs Stemming? Exact Functioning. Available at: https://medium.com/analytics-vidhya/lemmatization-vs-stemming-exact-functioning-b470a7db15db [Accessed: 20 April 2022].

Mallet: MAchine Learning for LanguagE Toolkit. [no date]. Available at: https://mimno.github.io/Mallet/index [Accessed: 24 March 2022].

Matplotlib 2021. Matplotlib — Visualization with Python. Available at: https://matplotlib.org/ [Accessed: 27 March 2022].

MonkeyLearn 2021. Sentiment Analysis: A Definitive Guide. Available at: https://monkeylearn.com/sentiment-analysis/ [Accessed: 22 March 2022].

Nkuna, J. 2020. What Twitter & Topic Mining reveal about the Beruit Explosion. Available at: https://medium.com/@juliansteam/what-twitter-topic-mining-reveal-about-the-beruit-explosion-23b669c53f48 [Accessed: 13 April 2022].

NLTK :: Natural Language Toolkit. 2022. Available at: https://www.nltk.org/ [Accessed: 24 March 2022].

pandas [no date]. pandas - Python Data Analysis Library. Available at: https://pandas.pydata.org/about/index.html [Accessed: 22 March 2022].

Pelgrim, R. 2021. Short-Text Topic Modeling: LDA vs GSDMM. Available at: https://towardsdatascience.com/short-text-topic-modeling-lda-vs-gsdmm-20f1db742e14 [Accessed: 1 May 2022].

Pickett, L. et al. 2021. Electric vehicles and infrastructure. Available at: https://researchbriefings.files.parliament.uk/documents/CBP-7480/CBP-7480.pdf [Accessed: 21 March 2022].

Prabhakaran, S. 2018. Lemmatization Approaches with Examples in Python. Available at: https://www.machinelearningplus.com/nlp/lemmatization-examples-python/ [Accessed: 20 April 2022].

Prabhakaran, S. 2018. Topic Modeling in Python with Gensim. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/ [Accessed: 24 March 2022].

Rabindranath, G. 2022. Topic Model Evaluation. Available at: https://highdemandskills.com/topic-model-evaluation/ [Accessed: 2 May 2022].

Rabindranath, G. 2022. Topic Modeling with LDA Explained: Applications and How It Works - HDS. Available at: https://highdemandskills.com/topic-modeling-intuitive/#h2-2 [Accessed: 20 April 2022].

Rude, B. 2022. Using Tweepy to analyze Twitter Data with Python. Available at: https://brittarude.github.io/blog/2021/08/01/Using-Tweepy-to-analyze-twitter-data [Accessed: 21 March 2022].

Sakellariou, A. 2022. $335 Million In Luxury Cars Lost After Cargo Ship Sinks In Atlantic Ocean. Available at: https://www.therichest.com/rich-powerful/millions-luxury-cars-lost-cargo-ship-sinks-atlantic/ [Accessed: 25 April 2022].

seaborn: statistical data visualization — seaborn 0.11.2 documentation. [no date]. Available at: https://seaborn.pydata.org/ [Accessed: 27 March 2022].

Seth, N. 2020. Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim. Available at: https://www.analyticsvidhya.com/blog/2021/06/part-3-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/ [Accessed: 4 May 2022].

Suresha, H. and Kumar Tiwari, K. 2021. Topic Modeling and Sentiment Analysis of Electric Vehicles of Twitter Data. Asian Journal of Research in Computer Science , pp. 13-29. doi: 10.9734/ajrcos/2021/v12i230278.

Thematic [no date]. Sentiment Analysis: Comprehensive Beginners Guide. Available at: https://getthematic.com/sentiment-analysis/#:~:text=Sentiment%20analysis%20uses%20machine%20learning,based%20and%20automated%20sentiment%20analysis [Accessed: 23 March 2022].

Tridens Technology 2022. EV charging infrastructure challenges and smart charging. Available at: https://tridenstechnology.com/ev-charging-infrastructure-challenges-and-smart-charging/ [Accessed: 26 April 2022].

Tutorialspoint [no date]. Gensim - Creating a Dictionary. Available at: https://www.tutorialspoint.com/gensim/gensim_creating_a_dictionary.htm [Accessed: 26 March 2022].

Tutorialspoint [no date]. Gensim - Creating LDA Mallet Model. Available at: https://www.tutorialspoint.com/gensim/gensim_creating_lda_mallet_model.htm [Accessed: 26 March 2022].

Tutorialspoint [no date]. Gensim Tutorial. Available at: https://www.tutorialspoint.com/gensim/index.htm [Accessed: 26 March 2022].

Ylä-Anttila, T. et al. 2021. Topic modeling for frame analysis: A study of media debates on climate change in India and USA. Global Media and Communication 18(1), pp. 91-112. doi: 10.1177/17427665211023984.