Individual Coursework Submission Form

Specialist Masters Programme

Surname: Xiao	**First Name: Chuqiao**
MSc in: Business Analytics	**Student ID number:**
Module Code: SMM799	
Module Title: Applied Research Project	
ARP tutor name: Philippe Blaettchen	**Submission Date: 01/09/2023**

Declaration:

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: **%**

Analysis on the Potential Evolution of Customer Perception towards Abarth 500e Electrification through Online Discussions

This report is submitted as part of the requirements for the award of the MSc in Business Analytics

Chuqiao Xiao

September 2023

Abstract:

Abarth is launching its latest car model 500e with new features. How the online discussion forms the meaning system that can influence the sales of product is critical for Abarth to further implement its strategy. By utilising the online comments from mainly 3 data sources, the online meaning system is represented and analysed with the assistance of NLP technologies. The compositions of customer perceptions on Abarth as well as its competitors are investigated, and the evolutions of online customer attitudes and sentiments are analysed. It's recognised that customers are focusing more on EV features and can be sensitive to EV prices in recent years. Some brands' strategies on developing intelligent systems can intrigue customers' interests, while customer perceptions on some granular brand values remain unclear.

Acknowledgement:

This year of MSc Business Analytics gave me much more surprising experiences and skills than my expectations. I would like to acknowledge the support given to me towards the completion of this dissertation.

First and foremost, I want to thank my supervisor, Philippe Blaettchen, for his support and professional advice throughout this project. He has provided us with constructive guidance and valuable feedback during the whole project.

Furthermore, I would like to thank all my colleagues who took the time to meet every week, organise our project progress and devote themselves to finalising this project. This research project would not have been possible without our collaboration.

Finally, special thanks to my family who supports both my academic and career decisions all the time. I wouldn't have accomplished the journey of learning and exploring without their company.

Table of Contents

List of Tables

List of Figures

1 Introduction

Electric vehicles (EV) are becoming popular all over the world in the past few years. The electric car sales exceeded 10 million in 2022 whilst the figure was around 1 million in 2017 (IEA, 2023). In particular, sales in the UK rose from 120,000 units in 2017 to 676,000 units in 2022 (Mintel, 2023). Wealthy households are the most important market (Mintel, 2023) and environmental awareness is one of the most vital values that motivates customers to purchase electric vehicles (Statista, 2023). However, the technology and facilities of electric vehicles are still under development, which can incur limitations when compared to traditional combustion engines. The limited charging accessibility is the greatest concern from the customers, and the relatively shorter range compared to the traditional combustion engine also dampens their demand (Statista, 2023).

Therefore, it's critical that a brand outlines the long-term strategy based on the potential opportunities and threats that come after a car model's electric transformation. In the age of the Internet, customer perceptions through online media are important and their behaviours can pose dramatic impacts on the sales of products (Micu, 2017). To gain a deep understanding of how the customers perceive the electrification of Abarth 500e, the online comments are collected and used as the text data to be analysed. Besides, close competitors that produce EV models with similar core features (e.g., compact design, sporty) are also the analysis targets of the research. It's believed that by applying contrast analysis, vital insights can be learned from Abarth's competitors' experiences. Mini, Peugeot and Volkswagen are identified as the close competitors by my colleague and Tesla is chosen as the EV industry benchmark.

The meaning systems associated with the Abarth 500e and other similar vehicles are pivotal to the interpretation of customer perceptions. Thus, Natural Language Processing (NLP) technologies are employed to analyse the text data collected by the web scraping pipeline. The analysing process is broken down into Exploratory Data Analysis (EDA), Topic Modelling, and Sentiment Analysis. Topics are summarised ('summarise' here indicates the process that researchers interpret topics based on both the relevant word distributions and original texts), then topic distributions before and after the launch of an EV model are compared. With the assistance of a pre-trained sentiment categorising model, each topic is assigned a sentiment score so that how the sentiment changes under each topic over time can be displayed on the plots. By analysing these results, how customers' perceptions changed can be depicted, and potential opportunities as well as threats can be predicted. The meaning system was interpreted based on the topic models as well as the top frequent words' distributions. With the aid of the sentiment analysis, how customers perceived Abarth 500e was summarised based on the topic-sentiment score models as well as the competitor analysis. How well the customer perceptions matched the core values of Abarth 500e was also examined.

On top of the analysis, recommendations on the strategic operations of the Abarth 500e will then be concluded. By leveraging these analysis results, Abarth would be able to address the threats with confidence and fully exploit the arising opportunities in the future.

2 Literature Review

Prior research and literature regarding the online discussion analysis with especially Topic Modelling would be thoroughly reviewed in this section. The following questions would lead the whole literature research:

1. How can topic models help interpret the meaning system founded on the online discussion?

2. Can the sentiment analysis help comprehend the meaning system?

3. How to interpret and explain the model results from a business perspective?

2.1 Meaning system based on online discussions

Online opinions are getting increasingly crucial with the development of online media, and the textual content posted is believed to have critical impacts on product demand (Lee, 2011). Moreover, with the development of the machine learning models, Lee (2011) proposed that the pattern change of consumer perceptions can be better captured by the automated algorithm which can possibly reveal valuable insights that the traditional methods (which are based on the predefined set of product attributes) might overlook. The automated clustering technology, for example, may highlight salient product attributes or identify unique submarkets. One of the popular technologies used is Topic Modelling. For instance, Puranam (2017) used topic models to analyse online reviews of New York City restaurants before and after a policy regarding showing calorie counts was carried out. The customer attitudes were proved to change in some patterns after the implementation. Marchetti (2020) used topic models to investigate how Netflix employees perceive corporate values and identified the mismatch between the positive values and the negative perceptions by some employees.

2.2 Topic model based on LDA

One simple analysis Puranam (2017) used to interpret the meaning system is by analysing the distribution of the most frequent words, based on a 'Bag of Words' method where he analysed the distributions of words after manually filtering out some meaningless words. On top of that, he analysed and compared word distributions before and after the implementation of the regularisation. However, it's implied that such a method doesn't provide any solid support for the interpretation and is also too vague and subjective. Topic model is believed to be a more reliable technology to help interpret online texts.

LDA topic model, in particular, was introduced by Blei (2003) and is applied to discover latent patterns of large bodies of text and then output a list of topics. It's based on statistical associations of words therefore the calculation is founded on the assumption that documents are a collection of words but no syntax (Blei, 2003). In the model result, the co-occurring words are put into the same clusters that indicate the underlying topics. The texts are composed of multiple topics in most situations (e.g., an online comment can indicate 35% of the topic related to the vehicle's appearance, 40% of the driving range, and 25% of the charging facilities), which implies a probabilistic interpretation of the meaning system (Puranam, 2017).

The initialising of a topic model relies on a predefined number of topics, and researchers have been looking for a way to tune this parameter. To identify the eventual model, Hannigan (2019) suggested that the fit of the model can be evaluated through either statistical method (e.g., coherence score) or validity (e.g., semantic validity). An alternative is that scholars in the fields of social sciences try to utilise both methods but also locate a balance between them (Hannigan, 2019).

2.3 Sentiment Analysis

Online customer sentiments can also help interpret the meaning system. For example, Marchetti (2020) used sentiment analysis to examine if the Netflix employees perceive the corporate values in a positive or negative way. Similarly, Puranam (2017) also mentioned 'sentiment' in the research. The increase in the amount of reviews was regarded as a positive attitude of the online opinion, but no sentiment classification methods were invoked. The sentiment is believed to be positively related to product sales as customer attitudes were proven to pose great impacts on product demand (Micu, 2017). The meaning system of online discussion can be interpreted more comprehensively with the help of both the topic model and sentiment analysis.

2.4 Result Interpretation

There are several ways to explain the results of topic model analysis. Referring to an analysis framework introduced by Puranam (2017), distributions of the most frequent words can be checked first. The topic distribution is the second stage where the occurrence of different topics can imply online customers' interests and how they shift. Additionally, the distribution of the volume of discussion is used as a supplement to see if more users are attracted or motivated to a certain topic. The last stage can be analysing the composition of a topic, because some underlying patterns may be revealed. For example, 'Potbelly' and 'Subway' were identified as close competitors as Subway was found in the top 10 words of the topic associated with Potbelly.

3 Methodology

The methodology section consists of two main phases: Data Preparation and NLP Analysis. Data sources and close competitors to be compared were analysed first and the NLP analysis was then implemented based on the scrapped text data. The competitors were already identified as Mini, Peugeot and Volkswagen by my colleagues, and Tesla was also selected as the industry benchmark to be analysed.

3.1 Data Preparation

3.1.1 Data Source Selection

To gain a comprehensive view of the online comments, 3 data sources are selected. SpeakEV is chosen as an online forum where EV aficionados gather, while Reddit is selected for its wide user range and active users. YouTube is also chosen for analysis since it's a video-based platform where valuable user behaviours might be detected. In particular, the analysing process of online comments from SpeakEV would be documented in detail in this report.

3.1.2 Data Collection

A scraping pipeline based on Selenium and Beautiful Soup was created to record the online comments and other information (e.g., dates posted). In particular, the scraping strategy on SpeakEV was based on the specific structural design of this website: There are forums of different brands and sub-forums that relate to different topics within each main forum. For each sub-forum, there are posts with a specific topic where users make comments and replies.

In accordance with the website's design, the scraping was executed as follows: To start with, enter the 'Forumns' page, find the target brand under the 'Vehicle Specific Forums' section. Redirect to the target brands' specific forums and record the link of all sub-forums. The strategy for scraping comments regarding Peugeot is different since there is no specific forum of this brand. As a result, 'Peugeot' was used as a keyword in the in-site search engine, and the output of results was recorded. One limitation here is that although the target object of this research is Abarth, there are only 89 comments containing the keyword 'Abarth' according to the in-site search results. The number of comments was considered to be too few to support the analysis, so the Fiat forum was then set as the close substitute data resource to Abarth.

An embedded iteration then scraped every comment and other associated information, including the user ID and date posted, etc. The text data of Fiat as well as close competitors are stored respectively.

3.1.3 Identifying Homogeneous Car Models and Dates of Release

The following table provided by my colleague displays the car models that have homogeneous features to the Abarth 500e and their dates of launch were recorded. These dates are important because they would be used later to identify which type of date influences the online discussion most.

Table 1: Launch dates of each brand

Brand	Announcement date	Media showcase	Release date
Abarth 500e	22 Nov 2022	22 Apr 2023	June 2023
new Mini Cooper SE	9 July 2019	11 Sept 2019	March 2020
Peugeot E208	25 Feb 2019	25 Feb 2019	Early 2020
VW E-Golf	14 Feb 2014		Summer 2014
Fiat New 500	4 March 2020		March 2021

3.1.4 Data Cleaning & Text Preprocessing

The design of the HTML web page of SpeakEV replies uses a nested foldable block. When a user replies to either the post or another comment, the replied content is also displayed and can be fully shown by clicking 'Click to expand…'. The scraping pipeline was only capable of noting down all the nested replies, which resulted in repetitive content within each reply. In the text cleaning phase, the 'Click to expand…' was used as the keyword to split text and the last part of the text slices (which is the reply itself excluding the previous comments) was saved.

The implementation of Topic Modelling relies on text tokenisation. By utilising the language model from spaCy, each text was transformed into a list of tokens that enabled topic models to summarise topics. Some of the frequent words that didn't contain useful information were excluded in the 2nd round of tokenisation (Singh, 2023). Nevertheless, it's expected that the model would capture the topics not only based on the individual tokens but also on the combined words which may create different meanings when two independent words are adjacent (e.g., electric vs. electric vehicle). Consequently, bi-gram was applied to the text data using the Phrases model from Gensim (Prabhakaran, 2022). The data was then ready for further analysis.

3.2 NLP Analysis

3.2.1 Simple EDA

The distributions of comments over time were visualised, and were compared to the table of model release dates. The lag between the date of release and the start of the most significant spike was identified. Time intervals from a spike started till it ended were also recorded.

The top 15 most frequent words were counted (Singh, 2023), and the word distribution within the spike was then compared with the word distribution of the whole time.

3.2.2 Topic Modelling with Tunning & Topic Interpreting

The range of topic numbers was set from 3 to 10, in order to prevent the topic clusters from being too sparse and granular to interpret. A pattern of overlapping also becomes increasingly

significant along with the increase of the number of topics, resulting in a higher difficulty to summarise topic meanings (Hannigan, 2019).

Then a hybrid tunning method was used to select the optimal hyperparameter. Firstly, the coherence scores of each topic number were calculated and visualised, according to which the optimal topic number was selected. The pyLDAvis also contributed to the decision of the topic number. The Inter-topic Distance Map was visualised via the multidimensional scaling through pyLDAvis, where topics were visualised in the form of bumbles within a plane-coordinate system. Referring to the distances between bumbles and their sizes, it can be inferred whether the current topic model performs well or not. When bumbles have similar sizes and fewer overlaps, they are recognised as being successfully identified (Marchetti, 2020).

Fiat Topic Model:

Based on the plot of coherence scores against the number of topics retained, the number of topics was set to 3. By checking the Intertopic Distance Map, it can be observed that there's no overlap between bubbles. Three clusters of similar sizes distribute with balance, suggesting that the model isolates topics successfully.

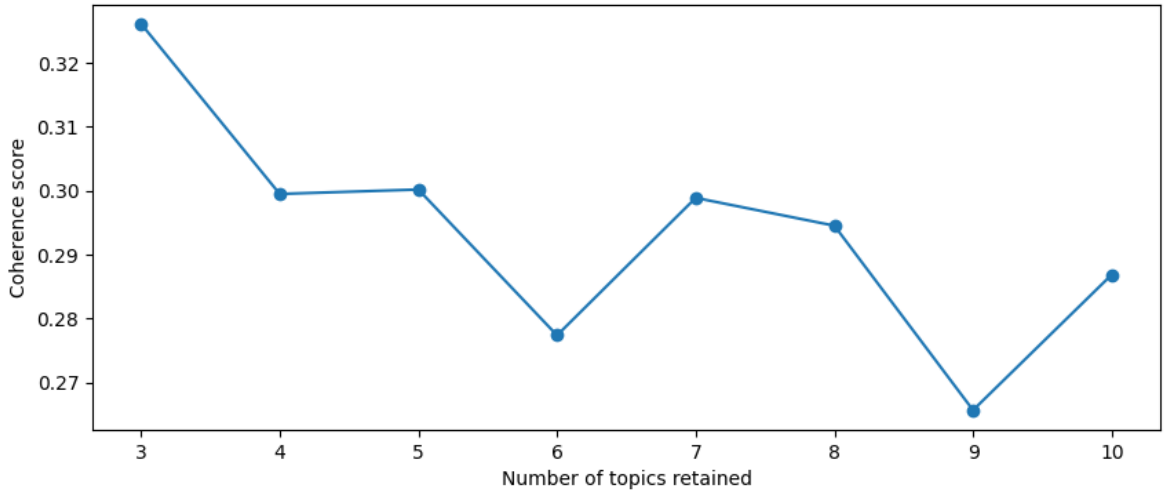


Figure 1: Coherence score against topic number of Fiat

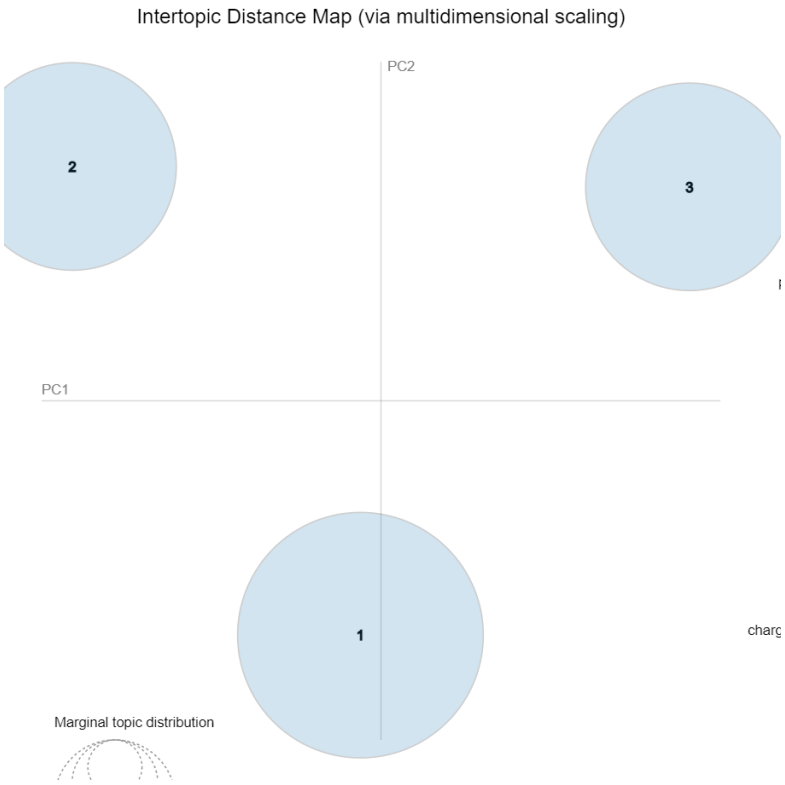


Figure 2: Intertopic Distance Map of Fiat

Peugeot Topic Model

Based on the plot of coherence scores against the number of topics retained, the number of topics was set to 3. By checking the Intertopic Distance Map, it can be observed that there's no overlap. Three clusters of similar sizes distribute with balance, suggesting that the model isolates topics successfully.

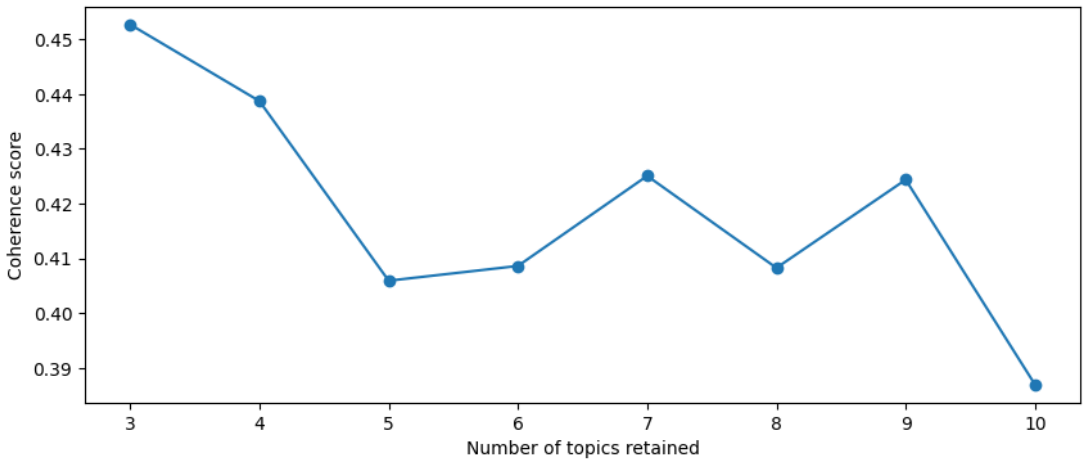


Figure 3: Coherence score against topic number of Peugeot

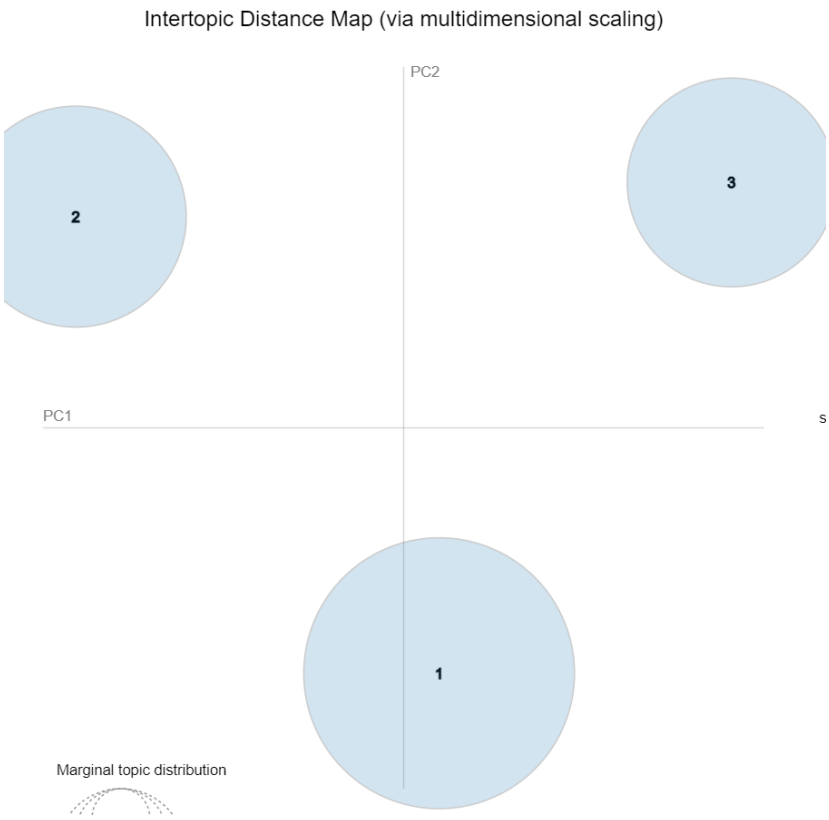


Figure 4: Intertopic Distance Map of Peugeot

Mini Topic Model:

Based on the plot of coherence scores against the number of topics retained, the number of topics was set to 5. By checking the Intertopic Distance Map, it can be observed that there are only a few overlaps between clusters 2, 3 and 4. Five clusters of similar sizes distribute with balance, suggesting that the model isolates topics successfully.

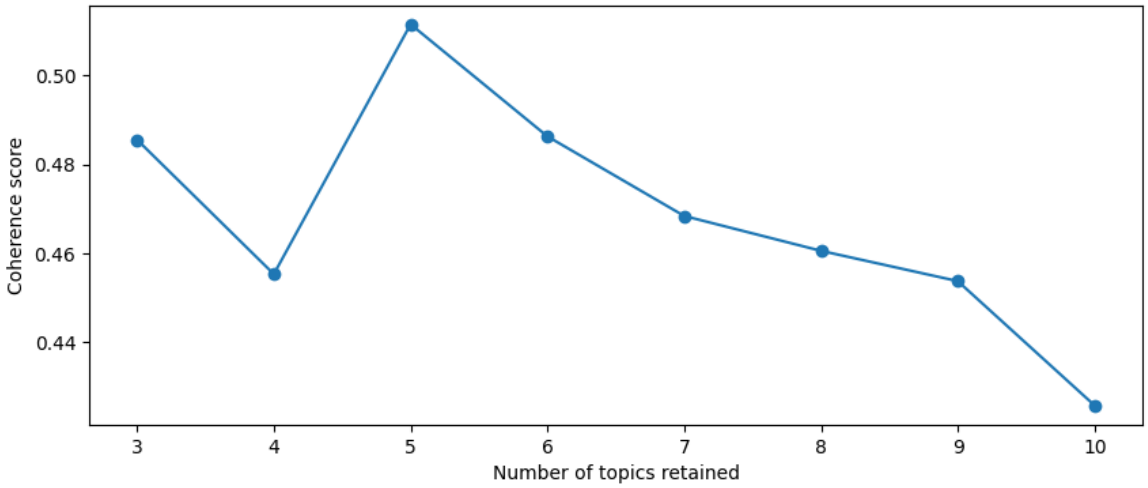


Figure 5: Coherence score against topic number of Mini

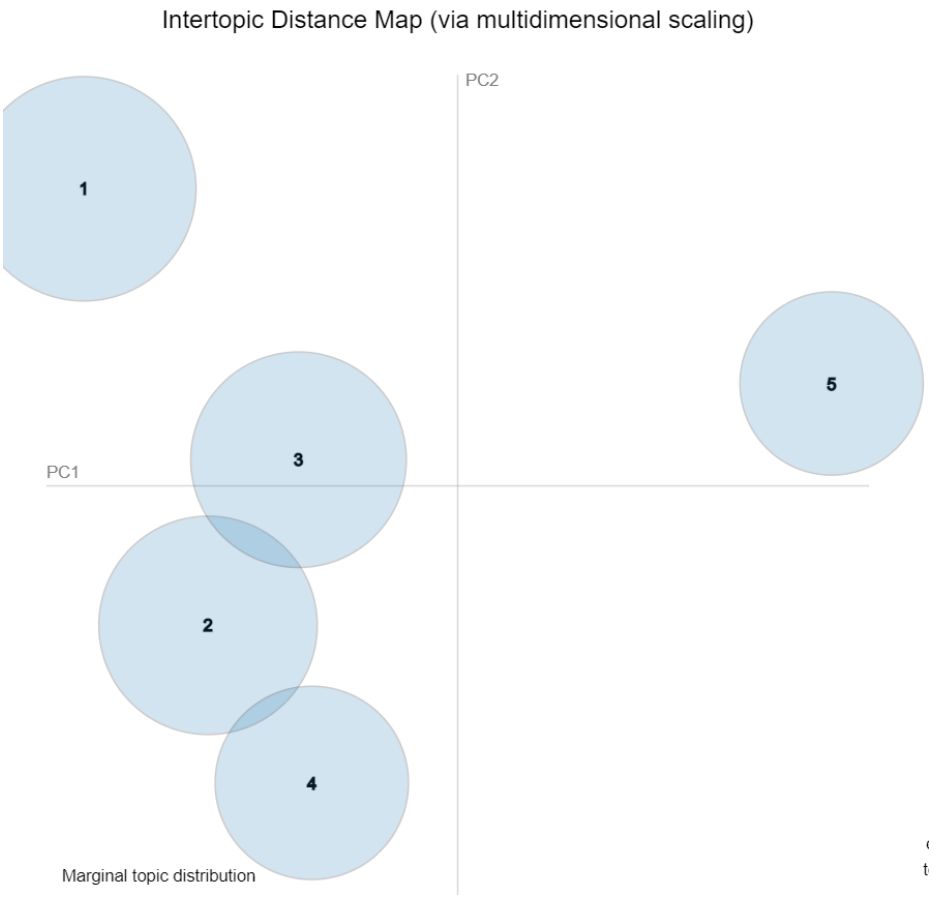


Figure 6: Intertopic Distance Map of Mini

Volkswagen Topic Model:

Based on the graphs below, the number of topics was set to 3.

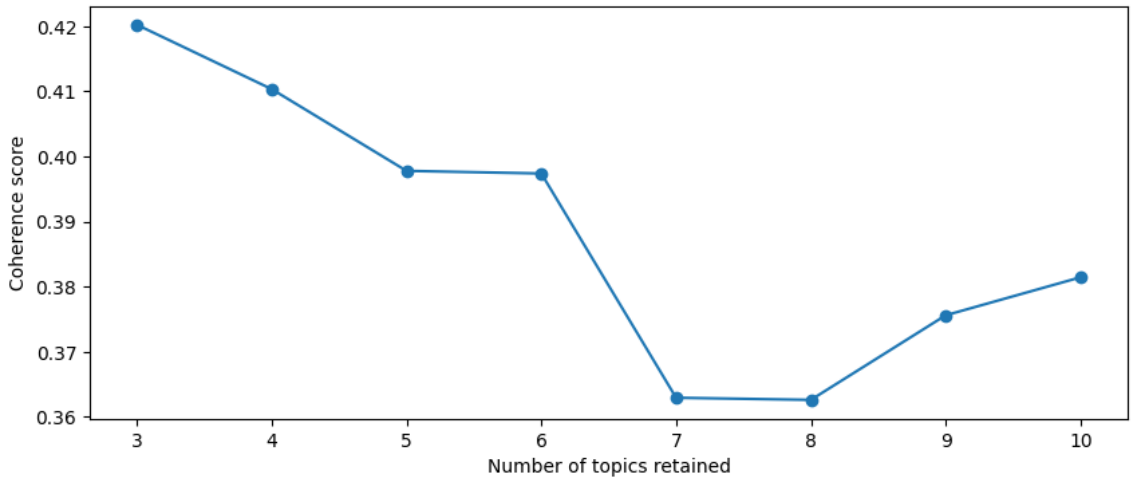


Figure 7: Coherence score against topic number of Volkswagen

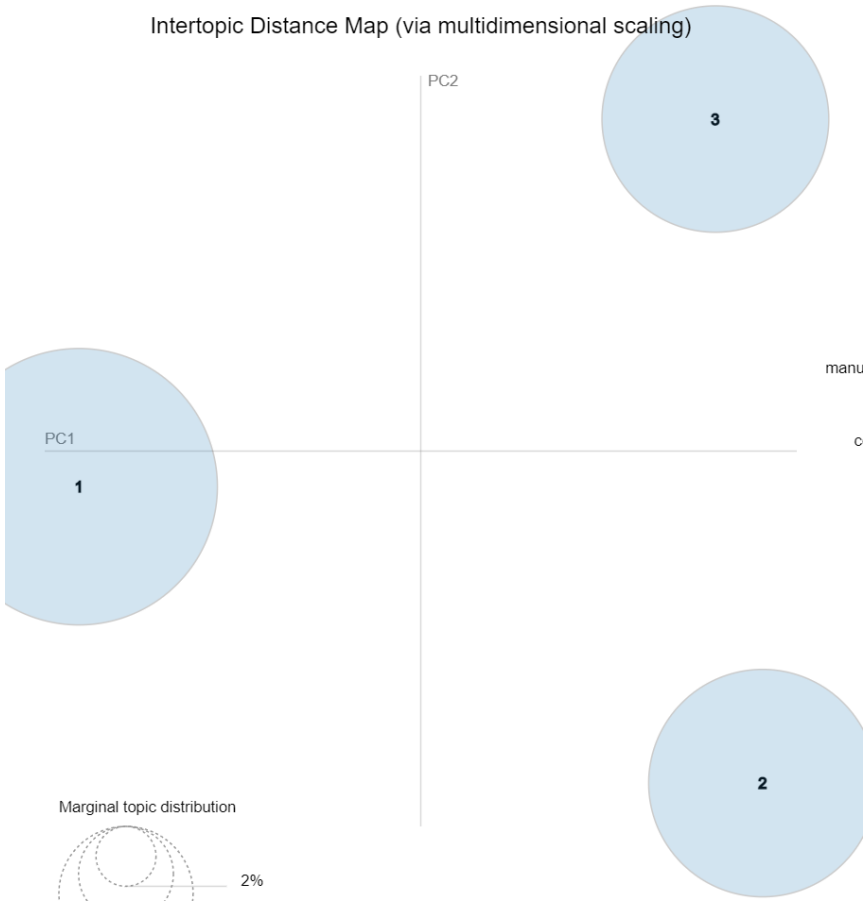


Figure 8: Intertopic Distance Map of Volkswagen

Tesla Topic Model:

Based on the graphs below, the number of topics was set to 3.

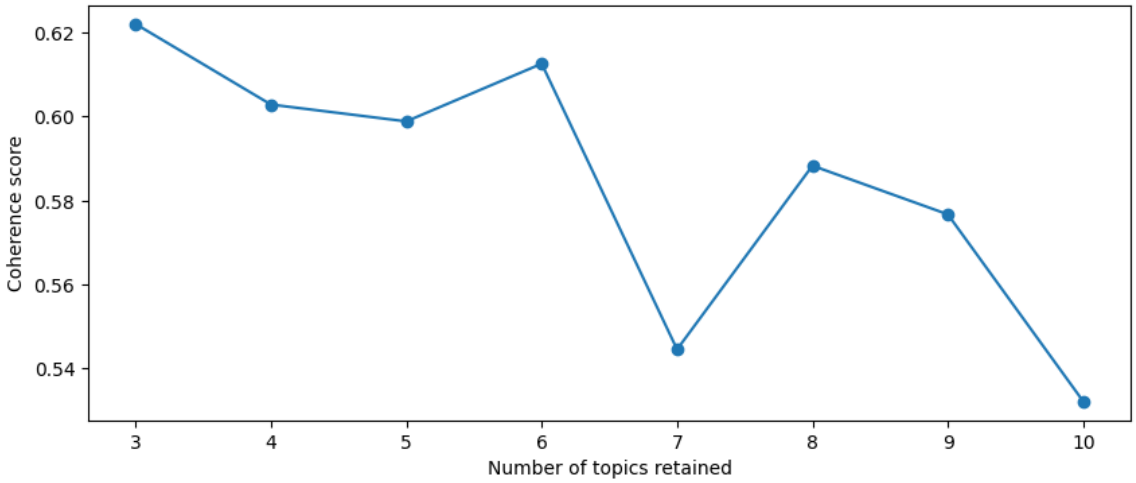


Figure 9: Coherence score against topic number of Tesla

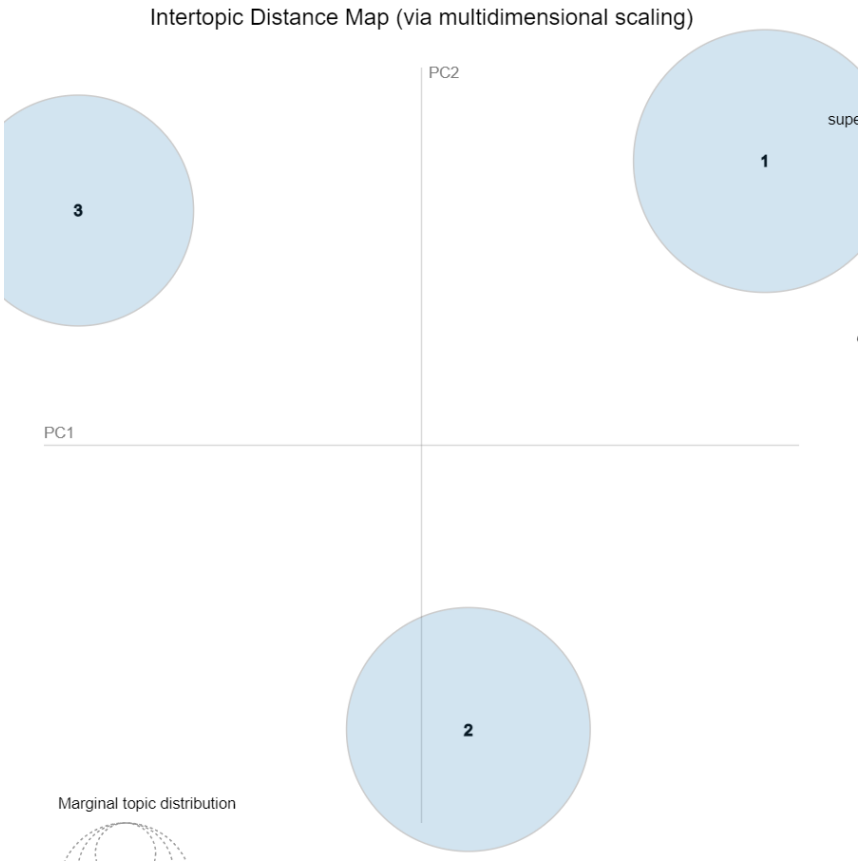


Figure 10: Intertopic Distance Map of Tesla

Founded on the previous filtering, the best fitting topic model was retained, and each text was assigned with one most relevant topic. A manual summarising process was then executed so that each topic was generalised into a brief description (e.g., Driving range & Battery).

3.2.3 Pre-trained Sentiment Analysis Model Selection

Moving on to the next stage, the sentiments associated with the specific topics are expected to be recognised. Given the lack of labelled data, two approaches can be applied: the sentiment classifier can be either trained by ourselves based on the manually labelled data, or from a Hugging Face pipeline of a pre-trained model. Although a self-trained (customised) model may better capture the jargon in the EV industry, it can cost too much time for the group to manually label the data and do the cleaning. Moreover, models from Hugging Face such as transformers are more advanced and with a larger scale of training data, which made them perform well on most tasks (Lokare, 2023). Consequently, the pre-trained model was eventually chosen as our categorising method.

For the concern of the efficiency of manual evaluation and simplicity, the model was expected to classify text into only 3 categories: positive, neutral, and negative. 500 random text samples were then selected and manually labelled by the group members, and 4 classifiers were examined in this report. The 'Twitter-Roberta-base-sentiment-latest' model from CardiffNLP (cardiffnlp, 2023b) outcompeted all other models including those from my other group members, reaching an accuracy of 0.712.

Table 2: Performances of pre-trained sentiment models

Data\Models	Model-Chuqiao	Model-Nadia	Model-Jia Meng	Model-Alvis
Data_500	Mdl 1: 0.47 Mdl 2: 0.588 Mdl 3: 0.708 Mdl 4: 0.712	N/A	Vader: 54% Model 2: 66.6%	Model: 69.31%
Selected Mdl:	**Mdl 4: 0.712**			

Due to the maximum length of the model being 512, texts were truncated to a length of 512 (cardiffnlp, 2022), each text was then assigned a sentiment label based on the selected model.

3.2.4 Topic Distributions & Sentiment-Topic Distributions

The result visualisation was conducted in the last section, where the frequency of topics under each brand was displayed first. A line graph was displayed for each brand where the change of topic frequencies over time was shown, and every topic was labelled with different colours.

On top of the topic visualisations, the distributions of comments across three distinct sentiment categories under each topic were displayed. The colours green, blue, and red were used to indicate positive, neutral, and negative respectively. Nevertheless, depicting graphs for each topic under every brand can be too redundant. In light of the opposite sentiment meanings that positive and negative hold, 3 distinct sentiment labels were quantified into numerical indicators where 1 implies a positive sentiment, 0 refers to the neutral sentiment and -1 indicates a negative sentiment. Sentiment points were then added up within each month so that the average sentiment score of each topic could be calculated. The simplified sentiment visualisation was then plotted where sentiments attached to topics were represented by the sentiment score.

4 Results and Analysis

This section seeks to explain and interpret the results of the methods mentioned previously in the NLP analysis phase from a business perspective. For each stage the analysis would be conducted after the result descriptions. Analysis regarding the meaning system, values perceptions and the identification of opportunities as well as threats are summarised at the last phase.

4.1 Simple EDA

The online comment distributions of each brand were visualised as below.

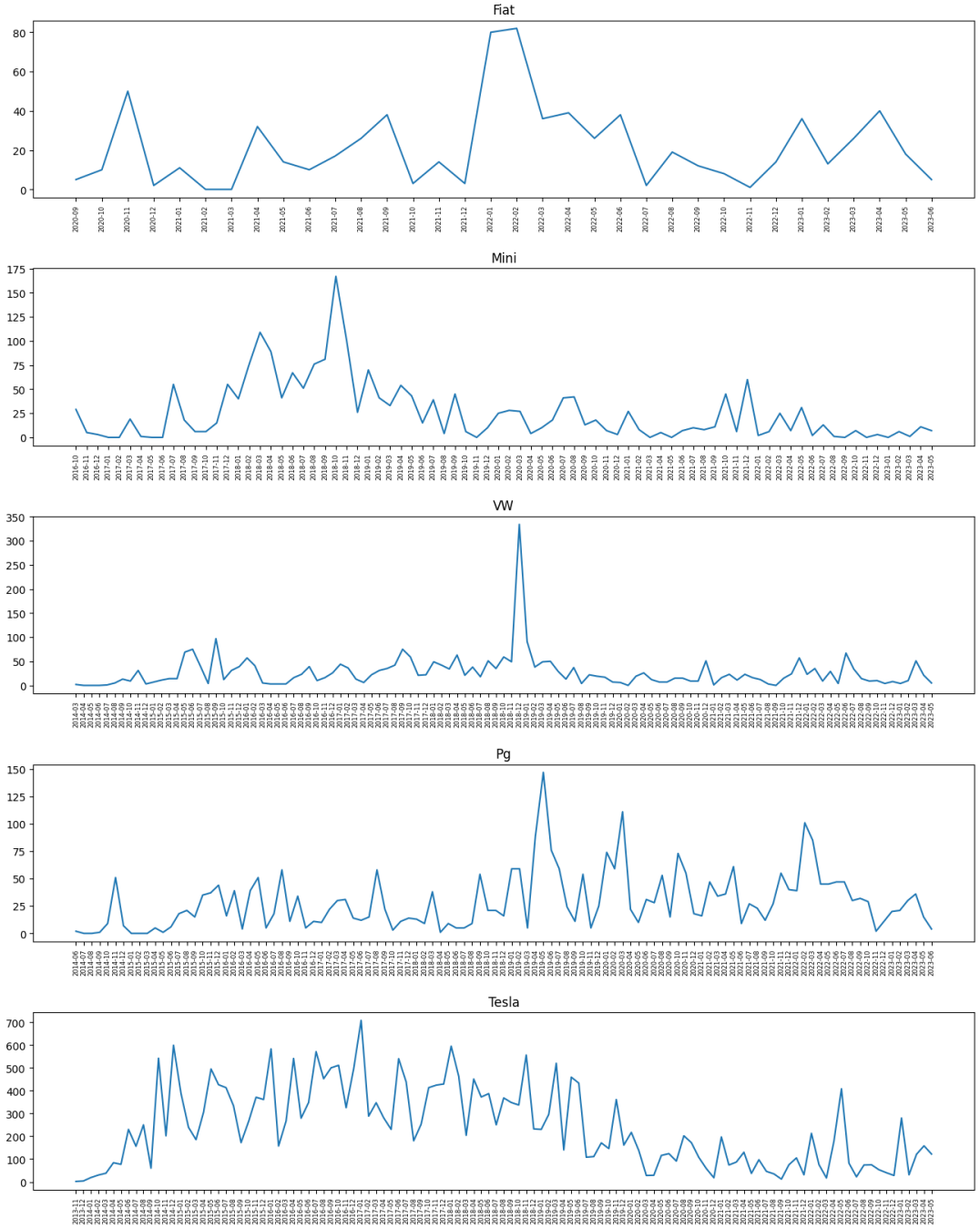


Figure 11: Data distributions of all brands

By analysing the graphs above and comparing these to the table of release dates, it's indicated that: Firstly, the forum of Fiat was created around 2 years ago and there are relatively smaller volumes of data and fewer active users. Some small spikes match the dates from the table of both Fiat new 500 and Abarth 500e, but patterns are not obvious. The most significant spike doesn't correspond to any dates listed. For Mini and VW (Volkswagen) as the close competitors, the dates of homogeneous car models don't match the spikes of online discussion. This might be attributed by that both brands are running on other car models which might gained more attention. By contrast, the distribution of Peugeot matches the 'Announcement date' perfectly, indicating that online discussion regarding Peugeot on SpeakEV might reveal insightful information that is more closely related to Abarth 500e. It's noticed that the online discussion reacted quickly to the date since the spike started to grow right after Peugeot E208 was announced.

In such a situation where Mini and VW's discussion may be less associated with the homogeneous car model, fewer patterns may be found by analysing data separated by the announcement date. The complementary solution was then implemented, where Peugeot was set as the main competitor to be analysed, while Mini, VW, and Tesla were identified as the other competitors. The 'dates' of those other competitors were defined as the start date of their most significant spike observed from the distribution. The dates and time intervals were recorded as below:

Table 3: Announcement dates and intervals of spikes

Brand	Announcement date	Interval (months)
Fiat	**20 Nov 2022**	**6**
Peugeot	**25 Feb 2019**	**7**
Mini	9 Jun 2018	7
Volkswagen	14 Oct 2018	5
Tesla	30 August 2019	7

The distributions of top 15 words within the spikes and through the whole time were then displayed. For Fiat, word distributions both relate to car range and battery. 'good' is a frequent token which indicates a possibly positive attitude.

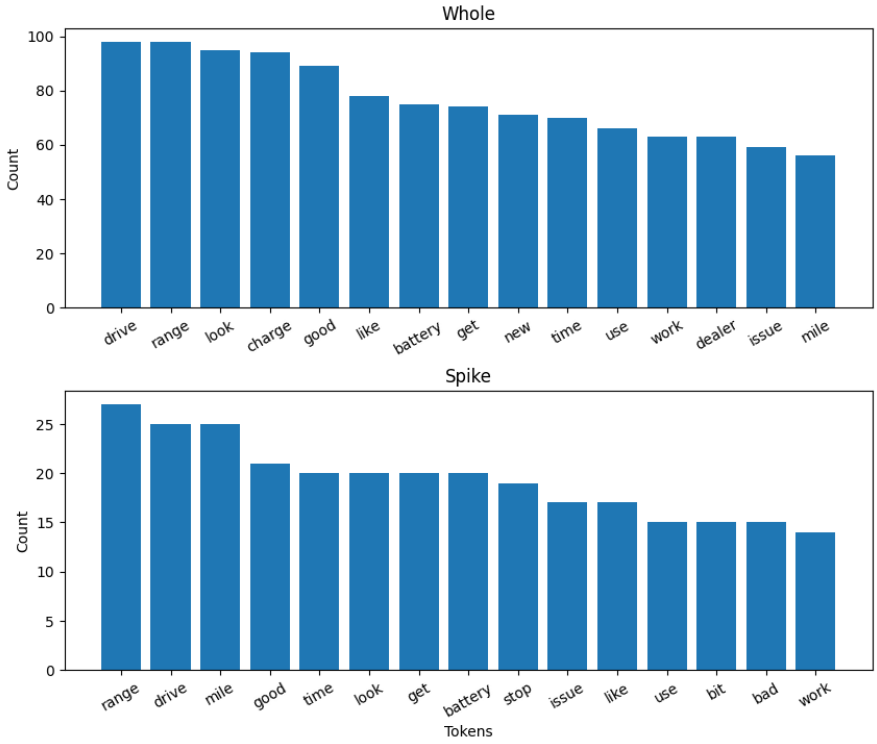


Figure 12: Distributions comparison of Fiat

In particular, word distributions of Peugeot are indicating useful insights. ‘price’ is dramatically more frequent (almost twice) than other tokens and tokens related to purchase such as ‘sell’, ‘buy’, ‘pricing’ are occurring, indicating the focus shift from general EV topics such as ‘Charging’ and ‘Range’ to the ‘Purchasing’.

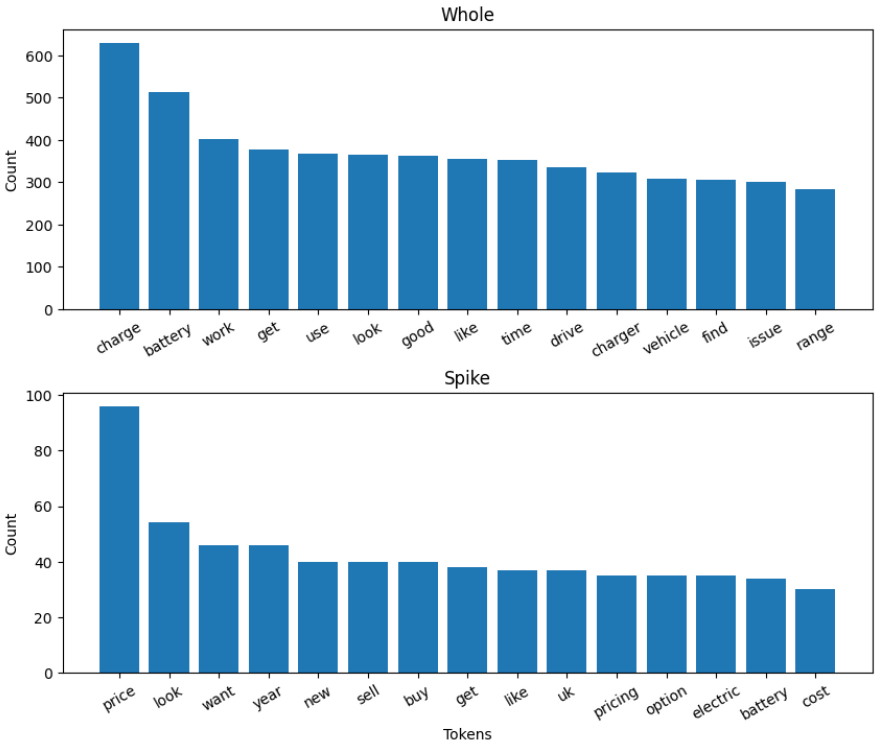


Figure 13: Distributions comparison of Peugeot

4.2 Topic Interpreting

4.2.1 Topic summarisation

In this phase, topic models of all brands were trained, and topics were summarised based on the relevant word lists, comment samples and human interpretation.

Fiat Topic Model

Topics were summarised below with the top 10 relevant tokens.

T1: General driving experience

range, drive, uk, dealer, cable, new, m, month, wheel, order

T2: Charging & Facilities

charge, charger, good, get, plug, try, work, seat, new, time

T3: Range & Battery

look, battery, like, range, drive, mile, good, get, app, set

Topic 1 is a bit vague for the interpreting which invokes topics regarding range, charging cable and wheel. Topics 2 and 3 are relatively consistent and associated with charging. Topic 2 focuses more on the facilities such as charger and plug whilst Topic 3 relates more to the battery and range. No specific patterns or clusters about 'sound', 'joy', 'sporty' were identified, which indicates a mismatch between Abarth's values and customer perceptions. This might be due to the limited volume of data.

Peugeot Topic Model

Topics were summarised below with the top 10 relevant tokens.

T1: Maintenance & Repair

battery, work, voltage, problem, try, ion, start, charge, issue, pin

T2: Charging (battery and range)

charge, battery, range, charger, drive, use, good, get, mile, work

T3: Purchase (price & dealer)

price, look, buy, sell, cell, year, pack, dealer, vehicle, tapatalk

Topic 1 is related to maintenance and repair, which is abnormal as an EV topic referring to all topics summarised. Considering the relatively smaller size of the Peugeot community, the very special topic might be related to a special post that intrigues other users' interests, or alternatively, this can be a special behaviour pattern of Peugeot fans. Topic 2 is associated with discussions about charging, battery and driving range. Topic 3 is related to car purchasing.

Mini Topic Model

Topics were summarised below with the top 10 relevant tokens.

T1: Charging (battery and charger)

charge, cable, get, charger, look, cost, work, time, use, drive

T2: Purchase and appearance

order, get, standard, pack, look, spec, option, good, add, black

T3: Purchase (dealer and range)

range, day, get, dealer, phev, mile, week, drive, order, bmw

T4: Service (dealer)

work, service, dealer, set, use, get, want, find, time, year

T5: General Discussion (performance and features)

drive, battery, mile, range, try, good, mpg, bit, brake, like

Topic 1 is a typical topic related to EV, which involves discussion about battery and charging facilities Topics 2, 3 and 4 are similar in terms of the category, when all of them are related to purchasing. Topics 2 and 3 are more related to the vehicle while topic 4 relates closer to dealer services. Topic 5 is more general and is associated with battery and driving range.

VW Topic Model

Topics were summarised below with the top 10 relevant tokens.

T1: Intelligent System & App

app, work, gte, phone, time, use, drive, try, net, like

T2: Charging & Dealer

charge, get, dealer, try, work, look, battery, charger, week, point

T3: Car Comparison

like, electric, tesla, year, big, sell, battery, diesel, model, well

Topic 1 is related to the usage of smartphone app which can be VW's feature. Topic 2 is related to charging and dealer service. In the 3rd Topic discussions are more about comparing VW with Tesla and other brands.

Tesla Topic Model

Topics were summarised below with the top 10 relevant tokens.

T1: General Discussion (Automation and others)

drive, like, time, thing, look, driver, good, road, work, way

T2: Purchase (price)

year, model, buy, model_s, like, price, cost, market, new, sell

T3: Charging & Battery

charge, supercharger, use, charger, time, work, battery, site, get, uk

Topic 1 is a general topic regarding automation and other car features. Topic 2 is related to purchasing and car prices. Topic 3 is associated with charging, facilities and battery.

4.2.2 Results Interpretations

Based on all the topics summarised, it can be noticed that two large categories of topics are very common and popular: Charging and Purchasing. The topic about charging involves discussion related to charging facilities, charging speed and convenience and battery which is also closely related to the driving range. The topic of purchasing involves discussions about the dealer, car prices (cost), buying and selling. General discussion can be related to other features of EV but none of them are as popular as the above 2.

Some repetitive patterns are frequently observed, for example, 'range' always co-occurs with 'battery', suggesting that customers discuss them together, so that these two features are actually mutually associated and influencing (Puranam, 2017). It's also noticed that for some leading brands such as Volkswagen and Tesla, intelligent systems (e.g., mobile phone app, automation) are frequently mentioned, indicating a feature to be developed that can potentially be a motivation of purchasing.

A drawback of these topic models might be that the numbers of topics are very few, resulting in the aggregated clusters of multiple sub-topics. For instance, battery and range were included in one topic. The aggregated clusters can lead to very general topics that may overlook patterns that are minor or subtle (Hannigan, 2019). However, the increase of topics may also influence the interpretability of clusters and there might be overlaps as well as variations in the size of clusters. To summarise, it's a balance between the granularity of topics and the readability of topics.

4.3 Topic Distributions & Sentiment-Topic Distributions

In this section, distributions of topics of Fiat, Peugeot would be mainly discussed, and distributions of Tesla would be analysed as a complementary benchmark to be investigated. Then the sentiment-topic distributions will be analysed comprehensively.

Fiat Distributions

The topic distribution over time is shown below. The frequency of Topic 1 (tagged as 0 in blue) was getting lower while Topic 2 and 3 (tagged as 1 in orange and 2 in green) was increasing after March 2022. The change of frequency indicates the interest shift from general discussion to the features related to charging, range, and battery.

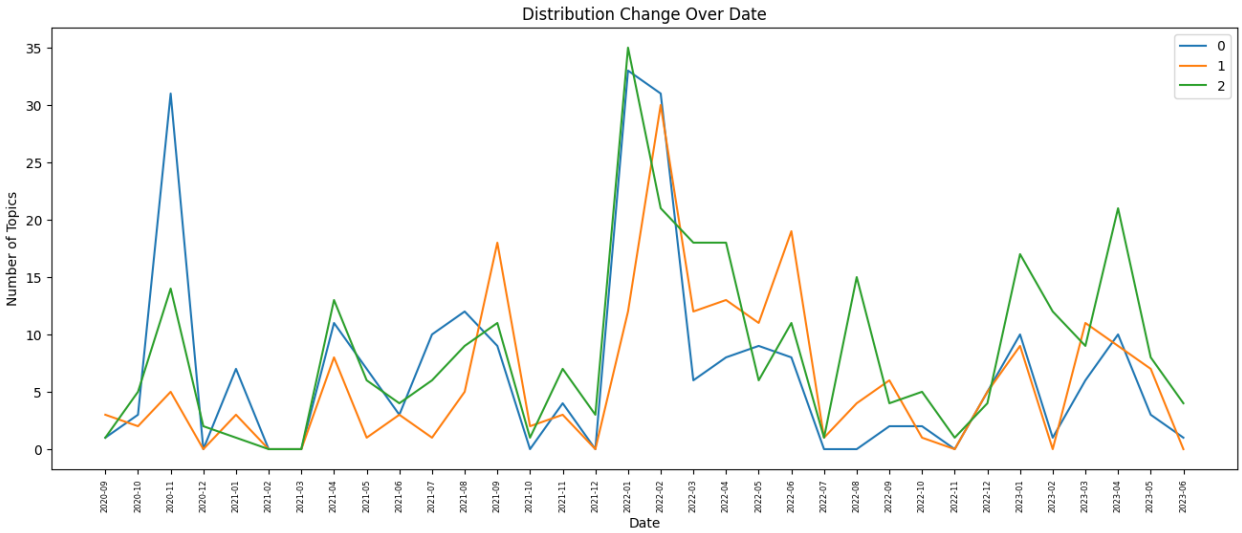
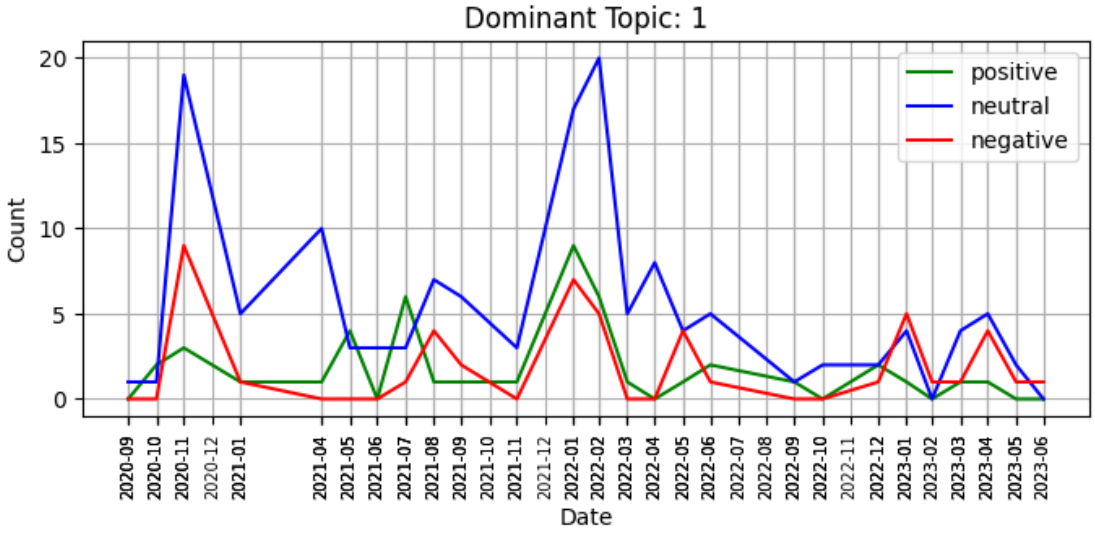


Figure 14: Topic distribution of Fiat

The sentiment distributions further help explain how online discussions perceive the product features. With Topic 1 and 3 having balanced distributions, a relatively obvious pattern can be identified in Topic 2, where the negative sentiments were high in the beginning of 2022 but more positive sentiments could be recognised at the beginning of 2023. Combining the topic summarisation in the last section, it's indicated that there's a positive evolution of online customer perceptions on Fiat's features in terms of charging & facilities.



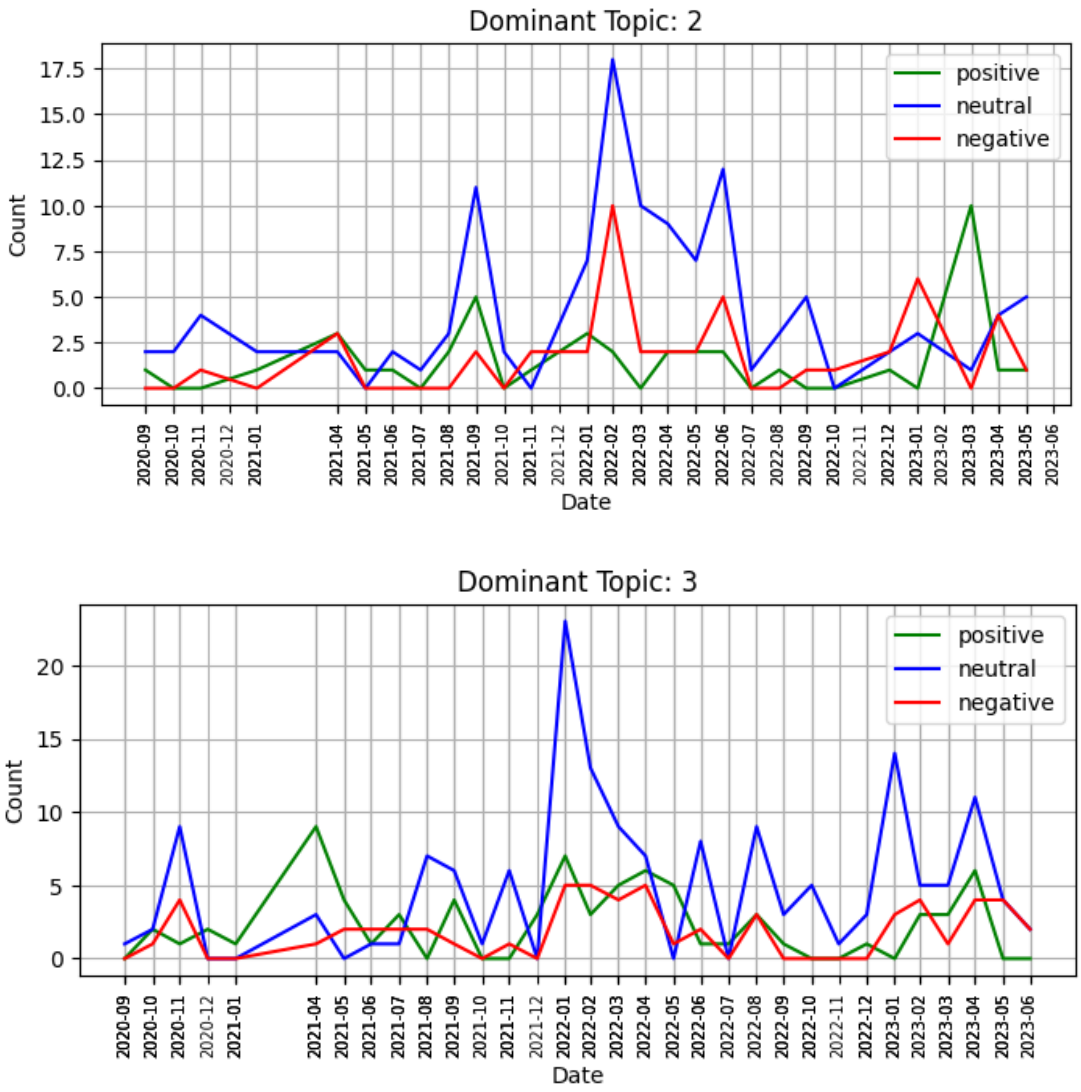


Figure 15: Sentiment distributions under each topic of Fiat

The sentiments were then added up and visualised in the form of sentiment scores below. It can be concluded that while the sentiments attached to Topic 1 went down at the beginning of 2023, positive sentiments associated with Topic 2 in 2023 had an uptrend after a drop at the end of 2022. The sentiments attached to Topic 3 remained to be above 0 (indicating positive) before 2023 but met a drop in May 2023. Overall, the sentiments attached to EV-related features are positively perceived, but customers may be losing interest in other car features and have been negatively perceived those features.

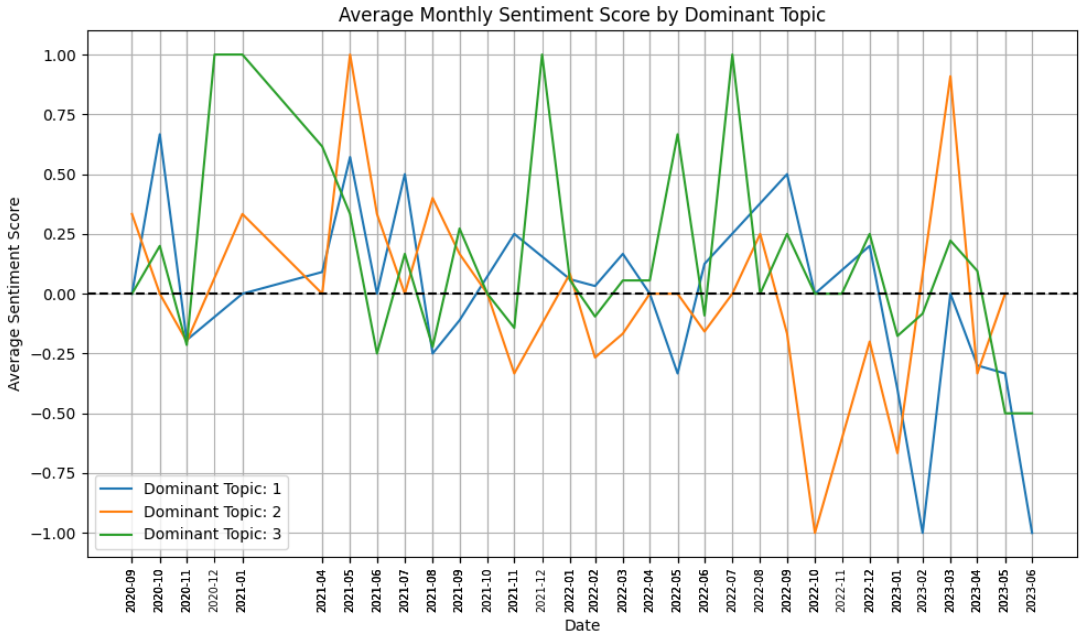


Figure 16: Sentiment-Topic distribution of Fiat

Peugeot Distributions

The topic distribution over time is shown below. The announcement date of Peugeot E208 was February 2019, and a high frequency of Topic 3 (tagged as 2 in green) can be recognised right after the date. Topic 2 (tagged as 1 in orange) remained popular throughout the whole observation. It can be interpreted from the topic distribution that Topic 3 regarding purchasing was frequently mentioned after the announcement date of Peugeot. This pattern recognised matches the analysis result from the simple EDA. Topic 2 kept being popular because it's related to the EV features which are the common topics for consumer discussion. Given the previous analysis on Topic 1, it's suggested that this topic may be severely influenced by some specific post content. Under such a situation, the meaning of Topic 1 couldn't be interpreted properly thus less attention was paid to Topic 1 in the following analysis.

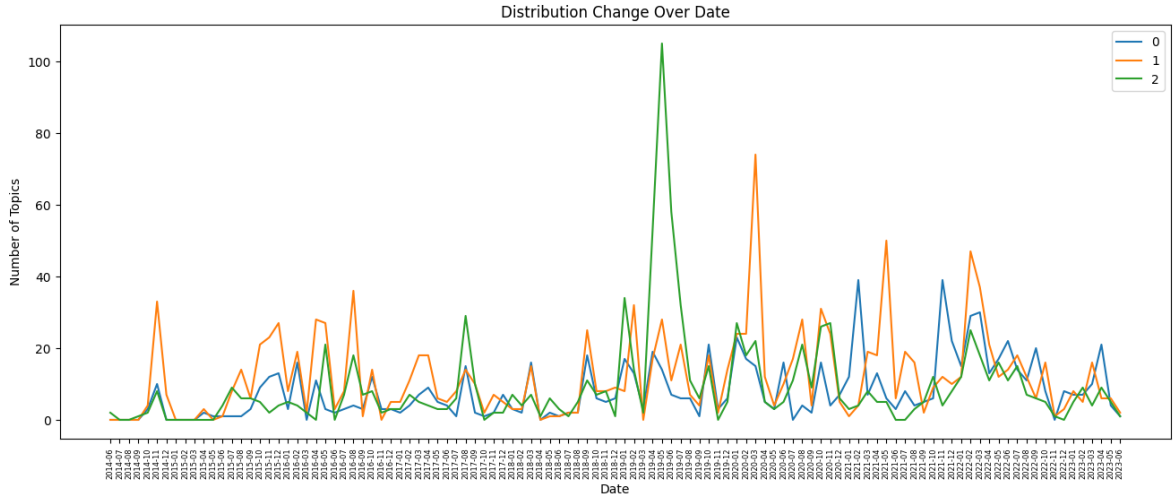


Figure 17: Topic distribution of Peugeot

The sentiment distributions further help explain how online discussions perceive the product features. Given the announcement date is Feb 2019, the analysis time range was narrowed to 2018-2023 so that the pattern changes can be better observed. There were slightly more positive sentiments attached to Topic 2 during the spike while more negative sentiments were associated with Topic 3. There were a few other spikes in April 2021, January 2022 where positive sentiments dominated online opinions. To summarise, discussions regarding car purchasing frequently occurred after the announcement of Peugeot E208 but with more negative sentiments, indicating a stronger focus but with negative customer perception and attitude. Customers might complain about the high cost of purchasing an electric vehicle at that time. On the opposite, topics regarding battery and charging remained positively perceived by customers, which indicated positive attitudes towards Peugeot's car features.

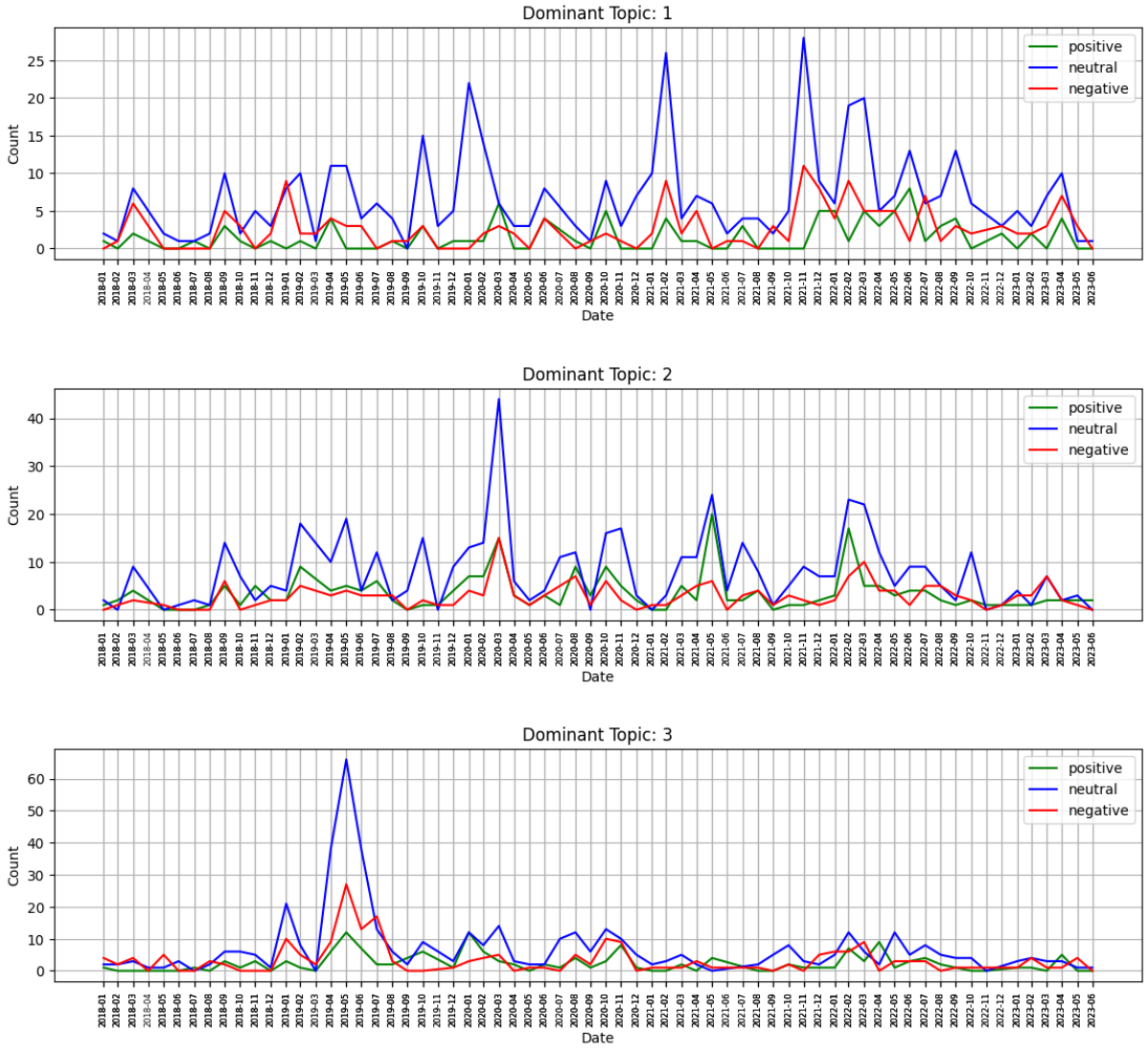


Figure 18: Sentiment distributions under each topic of Peugeot

The sentiments were then added up and visualised in the form of sentiment scores below. The distribution of sentiment scores of each topic matched the previous discussion, and it's proved that the topic related to price and car purchasing had been negatively perceived by online discussions overall. The positive sentiments regarding EV features were significantly

high in recent years, which might imply a positive trend in how online consumers perceive EV's functionalities.

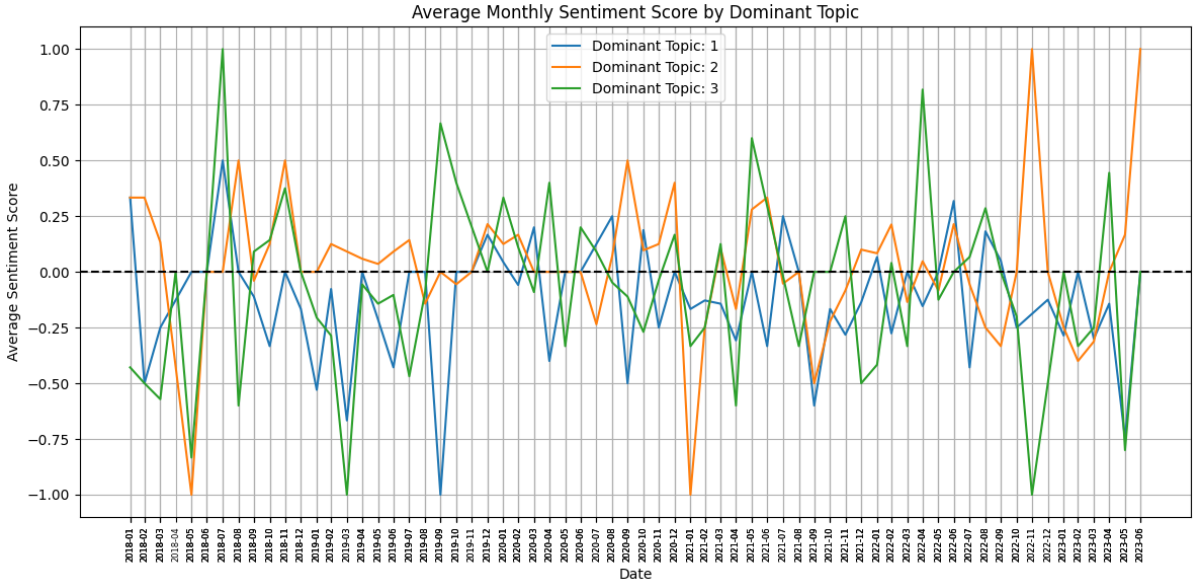


Figure 19: Sentiment-Topic distribution of Peugeot

Tesla Distributions

The topic distribution over time is shown below. The three topics seemed to have a similar volume of discussions. The overall popularity of Tesla dropped from 2020 to 2022, and has slightly increased since the beginning of 2022. Topic 2 and 3 (tagged as 1 in orange and 2 in green) were more popular compared to Topic 1 (tagged as 0 in blue) which is more general than the other two. The popularity of topics regarding purchasing and charging, battery aligns with the patterns that were concluded in the previous analysis.

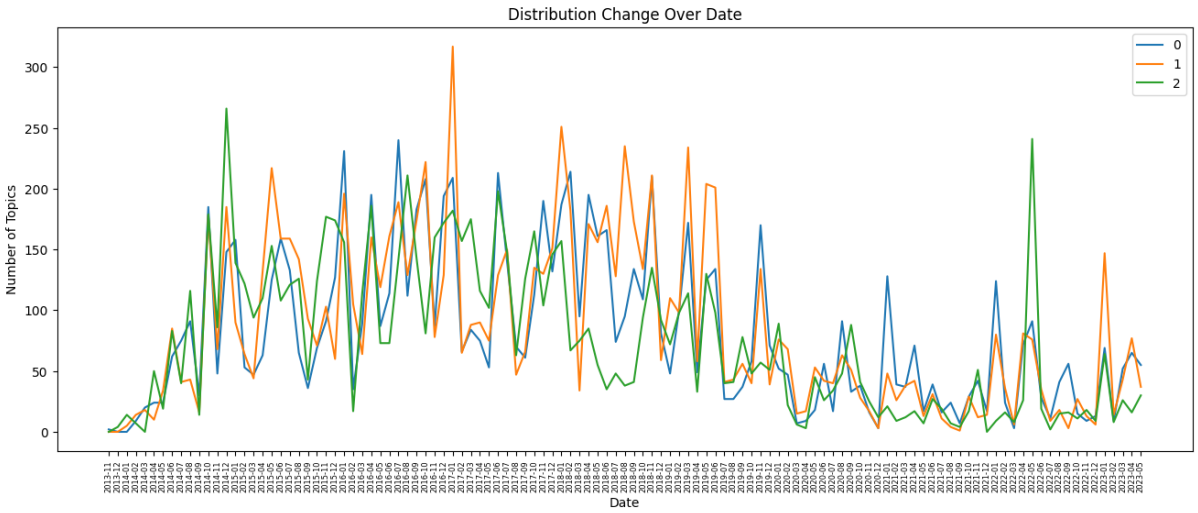


Figure 20: Topic distribution of Tesla

The sentiment distributions further help explain how online discussions perceive the product features. It seems that Tesla's online discussions have been perceiving its features in a negative way.

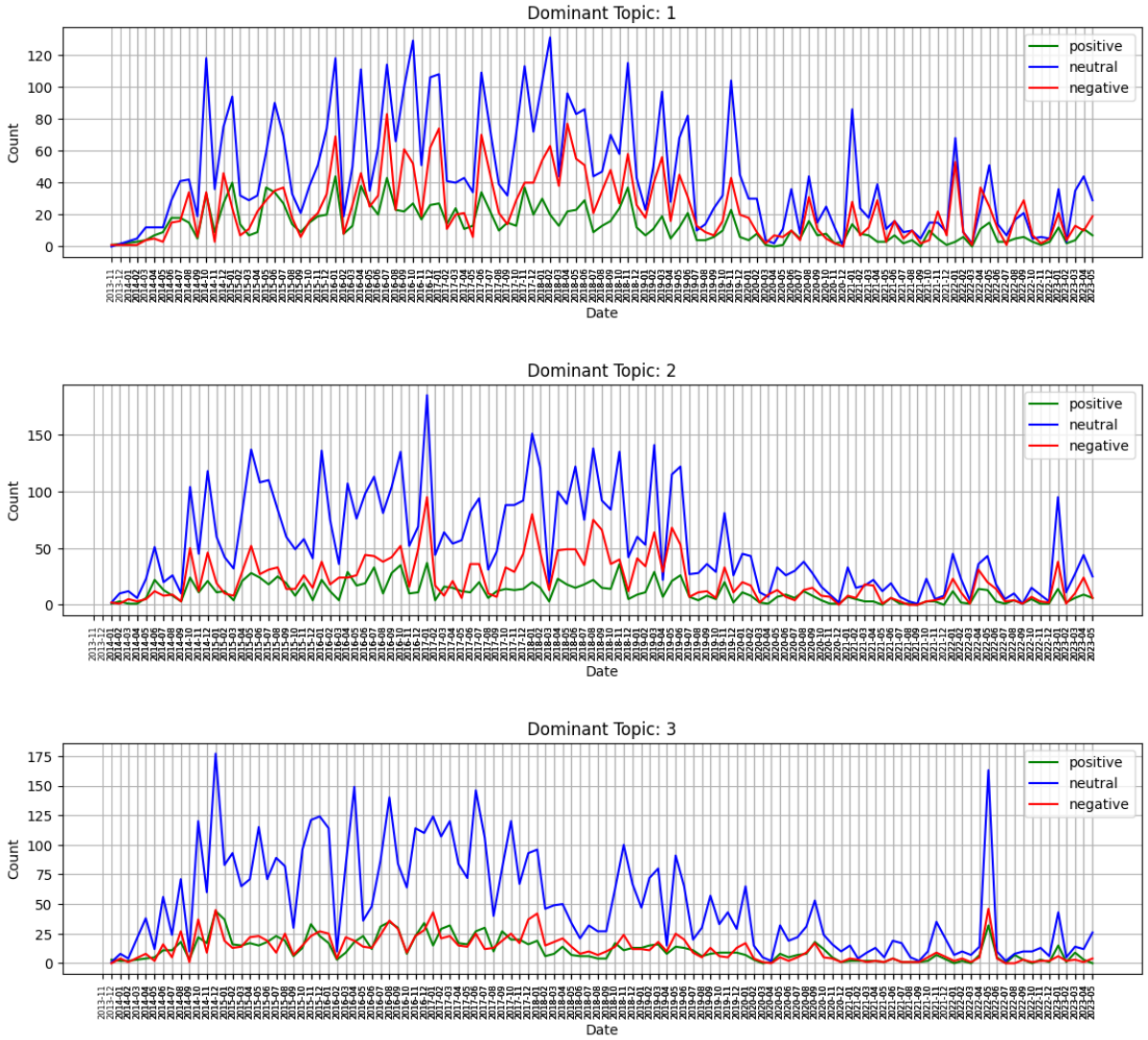


Figure 21: Sentiment distributions under each topic of Tesla

The sentiments were then added up and visualised in the form of sentiment scores below. It can be observed that sentiments attached to Topic 1 had a spike in 2021 but quickly fell below 0 (indicating negative sentiments). Topic 3 seems to be the only topic that was perceived in an overall positive attitude by customers, while Topic 2 had been perceived negatively. This pattern also matches the previous analysis that topics related to purchasing seemed to be perceived negatively while the EV features are more likely to be positively perceived.

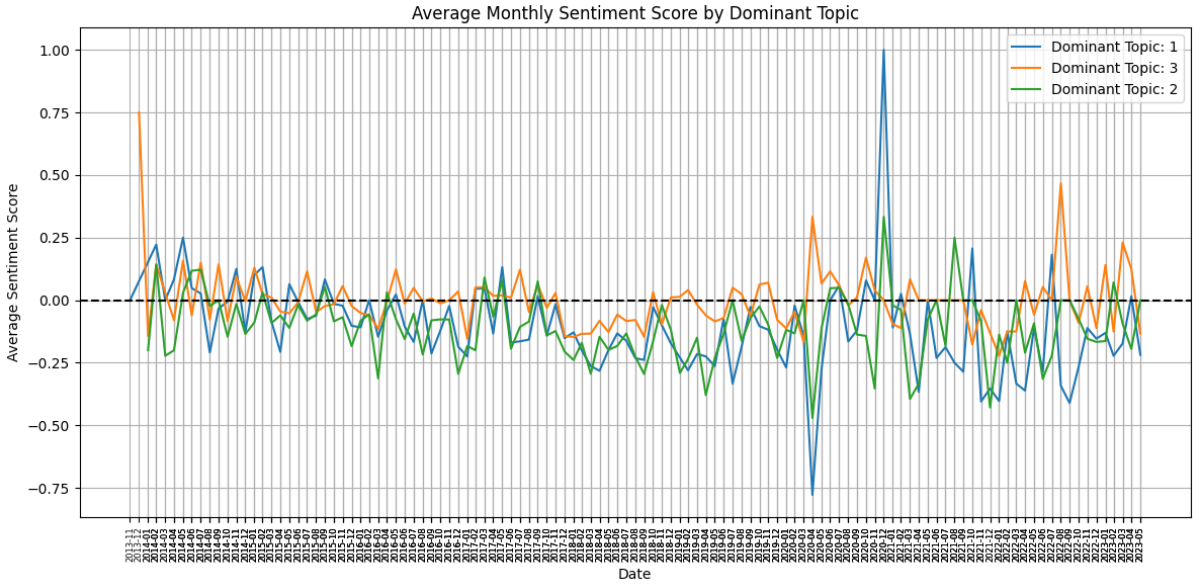


Figure 22: Sentiment-Topic distribution of Fiat

5 Conclusions & Recommendations

5.1 Analysis Conclusions

The meaning system associated with EV consists mainly of topics such as Charing System, Battery & Range, Purchasing & Price, and other general discussions according to the previous results. According to the topic distributions, customers are discussing increasingly more about battery, range and charging. This trend indicates a general enhanced interest in EV features and public awareness of the development of EV.

An overall shared pattern that can be identified is that topics related to purchasing and price are perceived negatively across all brands. This phenomenon can be explained by the research from Mintel (2023) which indicates that the cost and expense can be the main factor influencing consumers' (especially those at older ages) demands. Nevertheless, the positive perceptions of EV features such as charging and battery were recognised, regardless of these features are the top concerns from most customers (Statista, 2023) when is compared to traditional combustion engines. The rise in positive sentiments attached implies that the current EV performance, especially the performance of Fiat's EV was perceived positively by customers in recent years. This rising trend also indicates the weaker side of EV is gaining more positive attention, which can be the key for EV to outcompete traditional combustion cars.

It's also noticed that discussion related to purchasing increased dramatically after Peugeot's announcement, indicating that online consumers can be sensitive to the price which Abarth would like to pay attention to when releasing the future models. Tesla and VW are also exploring the intelligent systems such as mobile phone apps and automation, which can be a future opportunity for Abarth to exploit.

One limitation is that there is limited data related to Fiat Abarth on SpeakEV. It's a forum where only a small population of aficionados gather, so the discussions may not involve as many topics as what would be discussed on other platforms and the volume of data is also limited. According to Abarth 500e's value propositions, the core values that the band expect to deliver to customers are 'sound', 'joy', 'smart', 'sporty', etc. However, none of these values could be clearly isolated from the text data. No specific pattern changes after the announcement date can be recognised either. A possible explanation is that the topic model was tunned in a way that aggregates clusters too general, which results in a less granular topic distribution and overlooks patterns related to those values.

5.2 Recommendations

Based on the conclusions, some common patterns of EV discussion are summarised and threats as well as opportunities can be recognised.

According to the pattern concluded, customers are paying more attention to EV features such as battery and range, which are also their biggest concerns when comparing EV to traditional combustion cars (Statista, 2023). Abarth's high-level performance on the driving range should be proven so that customer confidence can be enhanced.

The pricing strategy should also be implemented carefully, since it's suggested that customers can be very sensitive to the cost according to Peugeot's sentiment-topic distribution.

A possible opportunity for Abarth is to develop an intelligent system such as a mobile phone app that helps control the car settings or car automation according to Tesla and VW's topic distributions.

Due to the limited data and the type of data source, no significant value perceptions were identified for any brands. For the Abarth's value delivery, the features regarding charging, range and price should be enhanced prior to other features.

6 References

Blei, D.M. *et al.* (2003) 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, 3(4/5), pp. 993–1022. Available at: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=bsu&AN=12323372&authtype=shib&site=ehost-live&scope=site (Accessed: 29 August 2023).

cardiffnlp (2022) *Cardiffnlp/twitter-roberta-base-sentiment-latest · Model Max length*, *cardiffnlp/twitter-roberta-base-sentiment-latest · Model max length*. Available at: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest/discussions/2 (Accessed: 29 August 2023).

cardiffnlp (2023a) *Cardiffnlp/twitter-roberta-base-sentiment · hugging face*, *cardiffnlp/twitter-roberta-base-sentiment · Hugging Face*. Available at: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment (Accessed: 29 August 2023).

cardiffnlp (2023b) *Cardiffnlp/twitter-roberta-base-sentiment-latest · hugging face*, *cardiffnlp/twitter-roberta-base-sentiment-latest · Hugging Face*. Available at: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest (Accessed: 29 August 2023).

Hannigan, T.R. *et al.* (2019) 'Topic modeling in management research: Rendering new theory from textual data', *Academy of Management Annals*, 13(2), pp. 586–632. doi:10.5465/annals.2017.0099.

IEA (2023), *Global EV Outlook 2023*, IEA, Paris. Available at: https://www.iea.org/reports/global-ev-outlook-2023 (Accessed: 22 August 2023)

Lee, T.Y. and BradLow, E.T. (2011) 'Automated Marketing Research Using Online Customer Reviews', *Journal of Marketing Research (JMR)*, 48(5), pp. 881–894. doi:10.1509/jmkr.48.5.881.

Lokare, G. (2023) *Effortless sentiment analysis with hugging face transformers: A beginner's guide*, *Medium*. Available at: https://medium.com/@lokaregns/effortless-sentiment-analysis-with-hugging-face-transformers-a-beginners-guide-359b0c8a1787 (Accessed: 29 August 2023).

Marchetti, A. and Puranam, P. (2020) 'Interpreting topic models using prototypical text: From "telling" to 'showing', *SSRN Electronic Journal* [Preprint]. doi:10.2139/ssrn.3717437.

Micu, A. *et al.* (2017) 'Analyzing user sentiment in social media: Implications for online marketing strategy', *Psychology & Marketing*, 34(12), pp. 1094–1100. doi:10.1002/mar.21049.

Mintel. (2021). Electric and Hybrid Cars- UK- 2023. Available at: https://reports.mintel.com/display/1101341/?fromSearch=%3Ffreetext%3Delectric%2520vehicle%26resultPosition%3D1 (Accessed: 22 August 2023).

nlptown (2023) *Nlptown/bert-base-multilingual-uncased-sentiment · hugging face*, *nlptown/bert-base-multilingual-uncased-sentiment · Hugging Face*. Available at: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment (Accessed: 29 August 2023).

Pascual, F. (2022) *Getting started with sentiment analysis using Python*, *Hugging Face – The AI community building the future.* Available at: https://huggingface.co/blog/sentiment-analysis-python (Accessed: 29 August 2023).

Prabhakaran, S. (2022) *Topic modeling in python with gensim*, *Machine Learning Plus*. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#18dominanttopicineachsentence (Accessed: 29 August 2023).

Puranam, D., Narayan, V. and Kadiyali, V. (2017) 'The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors', *Marketing Science*, 36(5), pp. 726–746. doi:10.1287/mksc.2017.1048.

sbcBI (2022) *SbcBI/sentiment_analysis_model · hugging face*, *sbcBI/sentiment_analysis_model · Hugging Face*. Available at: https://huggingface.co/sbcBI/sentiment_analysis_model (Accessed: 29 August 2023).

Singh, T. (2023) *Natural language processing with spacy in python*, *Real Python*. Available at: https://realpython.com/natural-language-processing-spacy-python/#the-doc-object-for-processed-text (Accessed: 29 August 2023).

Statista. (2023). Electric vehicles: A global overview. Available at: https://www-statista-com.eu1.proxy.openathens.net/study/134904/electric-vehicles-a-global-overview/ (Accessed: 22 August 2023)

7 Appendix

Designed Structure of Data Scrapped from SpeakEV

Post information:

Column Name	Column Description	Data Type
Index	Index of each post, but is also used as the post id in the comment dataset to indicate the hierarchical relationships between posts and comments	Int
Post_Title	Title of each post	String
Author	Author of each post, each user (author) is assigned with an id by SpeakEV	Int
Date	Date when the post was created	String
Post_Content	Content of the post	String
Comment_Number	Number of comments under one post	Float
Net_Likes	Likes of each post	Int
Views	Total views of the post	int

Comment information:

Column Name	Column Description	Data Type
Index	Index of each comment	Int
Comment_id	Unique id of each comment assigned by SpeakEV. Comment id would be used to show relationships between comments and replies	Int
Post_id	The post index stored in the Fiat_post_info.csv file. The post id indicates the hierarchical relationships between posts and comments	String
Author	Author of each comment, each user (author) is assigned with an id by SpeakEV	Int
Date	Date when the comment was created	String
Comment_Content	Content of the comment	String
Net_Likes	Likes of each comment	Int
Reply_id	If the comment is not replying to other comments, the value would be "None". If the comment is replying to other comments, the value would be the replied comment's id	String