Meaning System and Sentiments Attached to Abarth Pre- & Post-EV Launch

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

Alvis Cheung

1st September, 2023

Abstract

Abarth has just launched their first ever electric vehicle (EV), Abarth 500e. However, it does not match with their legacy in high-performance combustion engines and its roaring sound. In order to make Abarth's road to electrification smooth, online discussions were studied to understand Abarth fans' and audience's meaning system and sentiments attached to the brand before and after the EV launch, which helped to generate actionable strategies.

For such topic, we have two research focus: first, the meaning, which is to identify the major topics and keywords of the online discussion around Abarth; second, the sentiment, which is to understand the general emotion of the discussion, if it is positive, neutral, or negative.

Web Scraping was used to obtain the online discussion data; Topic Modelling and Sentiment Analysis were used for the analysis. We then compared the topics and sentiments across time to understand the evolution alongside three significant dates: Announcement, Public Showcase, and Market Release.

We found that people on Reddit usually talked about five topics of Abarth: 1) Maintenance/Engine, 2) General Usage, 3) Appearance, 4) Buying, and 5) Comparing. The sentiments of each topic were stable across time: Maintenance/Engine, General Usage, and Appearance discussions were mostly neutral; Buying and Comparing discussions were positive in general. We figured out that people were not satisfied with the general performance and maintenance of the traditional combustion engine Abarth, as there were more negative discussions than positive ones under topic Maintenance/Engine, despite being neutral in general. We also found out that people are doubtful towards the new Abarth 500e with the absence of a real engine and its iconic sound, as Maintenance/Engine and General Usage discussions were less positive and more negative after 500e's launch. Several recommendations were made addressing the opportunities and challenges.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Background

Electric vehicles (EV) have been in the market for years; however, they were never one of drivers' or manufacturers' options mainly due to the limitations on the battery: heavy, and low endurance (Standage, 2021). In recent years, Tesla's success in building a highly-rated EV (Consumer Reports, 2013), together with the pushing climate agenda, has brought people's attention back to EVs.

While the EV market is growing exponentially (International Energy Agency, 2023), the competition is fierce (Patton, 2022). If brands do not know how people think of their EVs, they would fail to take a share of this fast-growing market, and even lose their brand legacy.

Abarth is one the brands that are entering the EV market, by launching 'Abarth 500e', an electric version of their iconic Abarth 500. However, Abarth is well known for their high-performance combustion engines and their roaring sound (Stellantis, 2023), electrification of their cars would bring a huge challenge to this legacy. Even if the EV has been well-received by critics and enthusiasts, it is still anyone's guess how the audience of the fans will react.

1.2. Problem Formulation

In order to make Abarth's journey towards electrification smooth while preserving their legacy, understanding more about the brand's audience through appreciating online discussions can help provide actionable strategies.

The focus of analysis of this project would be in two areas: 1) meanings attached to the brand, and 2) sentiments around them. First, for the meaning, it is to identify the major topics and keywords of the online discussion around Abarth.

Second, for the sentiment, it is to understand if the discussion is positive, neutral, or negative.

With the comparison on these two areas across time, we can understand how Abarth's fans and audience perceive the brand and the electrification. Therefrom, we can identify opportunities and threats to the brand and its legacy, and thus provide actionable strategies.

1.3. Direction

Several techniques were utilised in addressing the research focus. First, Web Scraping was used to extract online discussion data from the internet. Next, Topic Modelling was used to identify the underlying contextual clusters of the discussion. Finally, Sentiment Analysis was used to understand the emotional tone of the discussion.

Topic Modelling has been shown to be effective in identifying patterns of discussions. Stokes et al. (2022), using Longitudinal Topic Modelling, found out that regarding COVID-19 the public generally talked about 1) public health measures, 2) daily life impact, and 3) sense of pandemic severity. Analysis within each topic also showed how people's priorities and concerns evolved over time. This echoes greatly to our study on the meaning and sentiment evolution in Abarth electrification: what are the main topics and meanings, and how they evolved over the course of electrification.

Sentiment Analysis has also been shown to work well in understanding actual emotions. In Philander and Zhong's (2016) study in hotel tweets sentiment, they successfully identified positive and negative events of the hotels using the highest and lowest sentiment scores. This shows that emotional shifts after an event can be identified using Sentiment Analysis models. Using such technique, it is possible to understand people's emotion towards Abarth electrification.

1.4. Report Focus

In this project, we worked as a group of four, including me, Jia Meng Low, Nadia Binti Muhammad Arif, and Shawn Xiao. This paper focuses mainly on my contribution to this project and my findings, on Reddit discussions about Abarth. All the related codes of the project are stored in GitHub[1], my contributions are under the branch 'Alvis'[2].

2. Methodology

Three techniques are used as the main tool for this study: 1) Web Scrapping, 2) Topic Modelling, and 3) Sentiment Analysis. We first collected online discussion data from several online discussion platforms via Web Scrapping. Then, we extracted the main topics of discussion with Topic Modelling. Finally, we examined the sentiments of the discussions using Sentiment Analysis. *Figure 1* below shows the whole process of natural language processing.

With all the topics and sentiments, we then compare these data along time, especially before and after three significant dates of the EV: 1) announcement, 2) public showcase, and 3) market release. This allows us to understand how the audience reacted to the new information they acquired from the brand.

In addition to comparing within the brand, we also compared the results across brands, where we can generate useful and accurate strategies with reference to Abarth's close competitors, namely Peugeot, Volkswagen, and Mini. Since Tesla is the EV industry leader (65.4% of total EV sales in 2022 (Rapier, 2022)), it is also included in the cross-brand comparison.

[1] https://github.com/XShawn1/Bayes_ARP_Abarth_1

[2] https://github.com/XShawn1/Bayes_ARP_Abarth_1/tree/Alvis

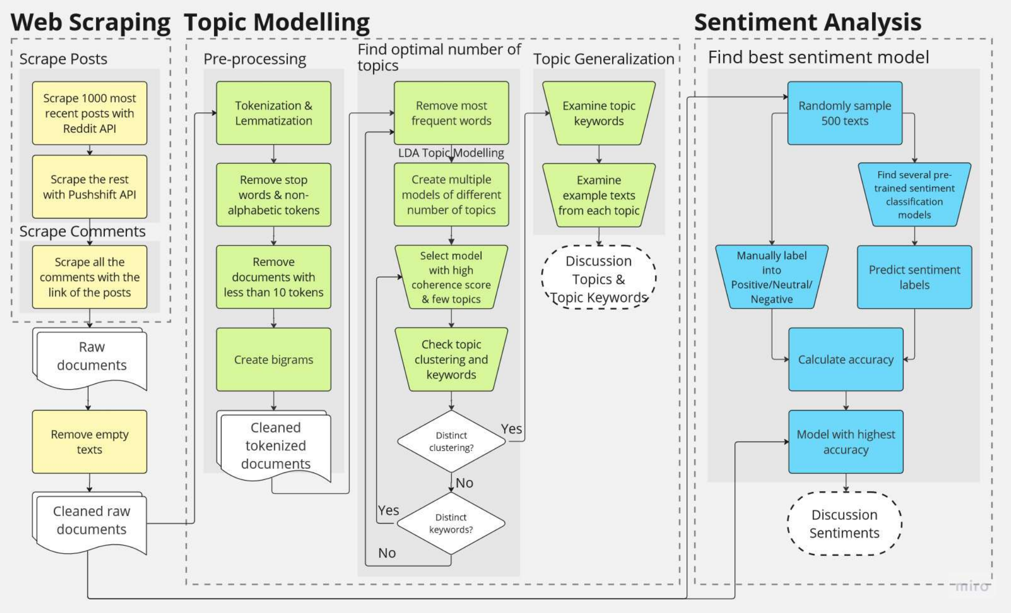
Web Scraping
Topic Modelling
Sentiment Analysis
Scrape Posts
Scrape 1000 most recent posts with Reddit API
Scrape the rest with Pushshift API
Scrape Comments
Scrape all the comments with the link of the posts
Raw documents
Remove empty texts
Cleaned raw documents
Pre-processing
Tokenization & Lemmatization
Remove stop words & non-alphabetic tokens
Remove documents with less than 10 tokens
Create bigrams
Cleaned tokenized documents
Find optimal number of topics
Remove most frequent words
LDA Topic Modelling
Create multiple models of different number of topics
Select model with high coherence score & few topics
Check topic clustering and keywords
Distinct clustering?
Yes
No
Distinct keywords?
Yes
No
Topic Generalization
Examine topic keywords
Examine example texts from each topic
Discussion Topics & Topic Keywords
Find best sentiment model
Randomly sample 500 texts
Find several pre-trained sentiment classification models
Manually label into Positive/Neutral/Negative
Predict sentiment labels
Calculate accuracy
Model with highest accuracy
Discussion Sentiments
miro

Figure 1: Flowchart of Natural Language Processing steps

2.1. Data Collection

We collected two types of data for our analysis: first, the online discussion data; second, the significant dates.

For the discussion data, we scraped from three different platforms to have a more comprehensive data: 1) Reddit, 2) SpeakEV, and 3) YouTube. Reddit is a forum that contains many kinds of discussions, which serves as a source of general discussion data in the project. SpeakEV is an electric vehicle-specific forum that most discussions are around EVs, which serves as a source of more specialized comments. Meanwhile, YouTube has numerous car review videos, which serves as a source of review comments.

For the significant dates, we simply did online research to identify the 3 dates: Announcement, Public Showcase, and Market Release. Announcement date is the date the brand announced they would launch their first EV; Public Showcase is when the brand showcased their newly developed EV to the public; Market Release is when the new EV was available to the market.

I am responsible for Reddit, including scrapping and analysis. The following parts of the Methodology section and the Result Analysis (*Section 3*) would focus on this.

2.1.1. API

The official Reddit API and the third-party Pushshift API[3] were used to scrape the posts and comments.

For Reddit API, I used PRAW[4] as the Python wrapper, which gives me access to the API in Python and has a limit of 1000 posts per request. So, I opted for Pushshift for access to much larger amount of data, however, the data is less updated as the Reddit API. At the end, I used Reddit API for the 1000 newest posts, and Pushshift for the rest/older that are available.

2.1.2. Data Source

As Reddit is composed of subreddits, which are community forums formed by users (Laukkonen, 2022), I directly scraped from specific Subreddits. First, I scraped the r/abarth subreddit, which is intuitive to go for. Then, in order to get greater amount of data, I scraped the r/fiat subreddit while searching for the keyword ‘abarth’, since Abarth is a division of the Fiat brand.

As posts data through both APIs only includes information of the post itself, the comments are not included, so I scraped for the comments separately through Reddit API. The 1000 limit does not matter as none of the posts has more than 1000 comments. I used the posts’ link scraped in the previous step as the input to get all the information of all the comments of the posts.

The most important information scraped was the texts and datetime of the posts and comments. Other info was scraped as well but did not used much, including the author, number of comments, score, id, etc (*Table 3-5 in Appendix*).

[3] https://reddit-api.readthedocs.io/en/latest/
[4] https://praw.readthedocs.io/en/stable/#getting-started

2.1.3. Timeframe

There were two rounds of scraping: mid-May and early-August 2023, as the project lasted for several months, new data should be incorporated into the analysis. The timeframe for the whole dataset spans from late-June 2013 to early-August 2023. However, data before 2017 was not used in the analysis. Reasons are twofold: 1) it is too old to be relevant to the electrification, 2) the data size is too small to be comparable and accurate.

2.1.4. Competitor Brands

Initially, we were targeting a different set of competing brands for cross-brand analysis, data were also scraped accordingly. However, the time we realized those were not Abarth's close competitors, Pushshift API became not available anymore with the update in Reddit API terms in April 2023[5]. With Reddit's official API allowing users to scrape 1000 posts only, it is not possible to scrape data for the close competitors for a comprehensive analysis. As a result, my scraping and analysis was on Abarth on Reddit only.

2.2. Topic Modelling

Topic Modelling allowed us to segregate the Reddit discussions into clusters of similar words, this helped us understand Abarth audience's focus and meaning to the brand. We first applied cleaning and tokenization on the text data, then utilised Latent Dirichlet Allocation (LDA) model to find out the hidden topics.

2.2.1. Preprocessing

The basic working units for LDA are 'tokens', which are individual words or punctuations. The tokens would be used to perform clustering and calculate the word-to-topic & document-to-topic probabilities (*Figure 2*). So, we had to

[5] https://www.reddit.com/r/reddit/comments/12qwagm/an_update_regarding_reddits_api/, https://www.reddit.com/r/modnews/comments/134tjpe/reddit_data_api_update_changes_to_pushshift_access/

tokenize each piece of text (document) before feeding it into the model. SpaCy is used as the Python library to do the preprocessing.

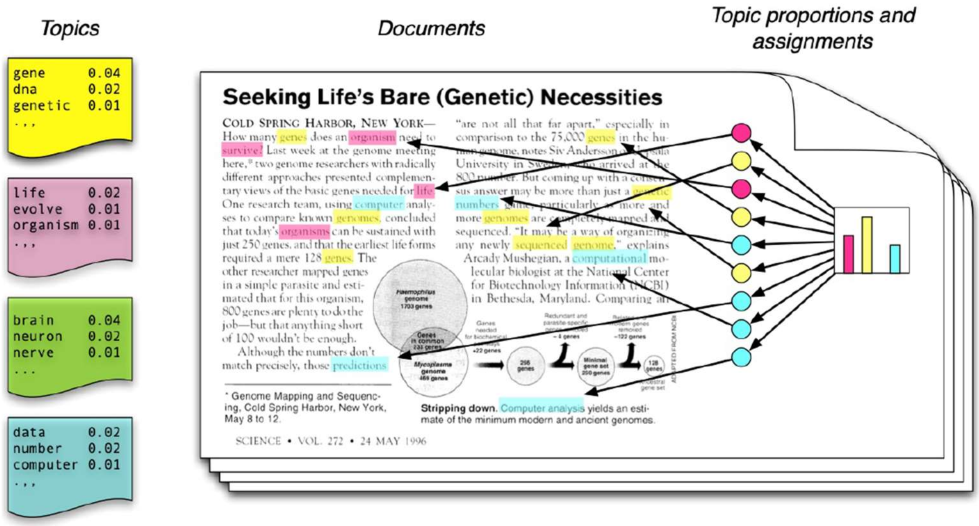

Figure 2: Popular image explaining LDA Topic Modelling (Bansal, 2016)

First, we combined the posts and comments dataset into one, and cleaned it up by removing any intuitively useless entries:

- '[deleted]': the author and his/her posts/comments were removed,
- '[removed]': the post/comment was removed,
- empty entries: the post/comment does not contain any text.

Then, each text was tokenized and lemmatized, generating lists of individual words of its root form (i.e. 'impress', 'impressive' and 'impressed' can all be reduced to 'impress'). With the lists of tokens, we then filtered out the stop words (words that 'does not add much meaning to a sentence' (Teja, 2020)), and kept only alphabetic tokens, since stop words and non-alphabetic tokens do not help us understand the context of the document.

Next, documents with less than 10 tokens were removed. A text that contains only few words would make clustering a hard and inaccurate job. Even if we can accurately assign them topics, there are not enough words to examine their sentiments.

We then created bigrams (combinations of two words) from each of token lists. It allowed us to take word combinations into account, e.g., United States, electric vehicle.

The final step of preprocessing was removing the most frequent words. Since those words would appear in many of the documents, removing them can prevent them from dominating all the topics, making the topics indistinguishable.

2.2.2. Number of Topics

After having a clean tokenized dataset, we proceeded to Topic Modelling. We utilized LDA to generate multiple models with a range of number of topics, using Gensim library.

Then, we examined how interpretable the topics are by calculating the coherence score for each model. Coherence score measures the similarity of words in each topic in the model (Zvornicanin, 2021). In Gensim, the higher the score, the easier the interpreting.

Alongside with the coherence score, we examined the models with the intertopic distance map using pyLDAvis library. It shows the distance between topics along 2 dimensions, which are reduced from the dataset. It helped to see if the topics are distinct enough.

With the coherence scores and visualizations, we can determine the optimal number of topics.

2.2.3. Topics Generalization

The output of the LDA model includes the doc-to-topic probabilities (likelihood of the document belonging to the topic) & the word-to-topic probabilities. We labelled the document with the topic of the highest doc-to-topic probability.

Then, we listed out the raw text of the highest probabilities for each topic. Together with the word-to-topic probabilities, we can understand the contextual meaning of the topics and can label the topics.

2.3. Sentiment Analysis

Sentiment Analysis combining with Topic Modelling allows us to understand the sentiment of the discussions for each topic. A simple text classification approach was adopted over a complex sentiment score, because distribution of labels can already show the general change in sentiment, while it is hard to determine if the change in sentiment score is significant or not.

2.3.1. Model Selection

We decided to use a pre-trained model directly instead of fine-tuning it with our own data and labels, since manual labelling would be time-consuming.

We selected several models to find the best for our dataset. To evaluate the model accuracy, we randomly sampled 500 texts from our dataset and manually labelled them into either Positive, Neutral or Negative, and compared them with the labels predicted by the models. We then chose the model with the highest accuracy, which is the 'Twitter-roBERTa-base for Sentiment Analysis - UPDATED (2022)[6]'.

2.3.2. Sentiment Classification

With the selected model, we then used all the raw texts as input to generate the sentiment labels.

[6] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

3. Results & Analysis

In addressing the meaning system attached to Abarth, using Topic Modelling techniques, we have identified the major topics people are discussing and the keywords people usually mentioned in Reddit.

While for the sentiments, we made use of Sentiment Analysis model to understand the general feeling of the Reddit users on Abarth.

We compared the meaning system and sentiment before and after Abarth's EV launch, by analysing the distributions of the topics, keywords, and sentiments before and after the significant dates.

3.1. Significant Dates

According to our online research, the Announcement, Public Showcase, and Release date for the new Abarth 500e are 22nd Nov 2022, 20th Apr 2023[7] and Jun 2023 respectively. For analysis, I set the Release date as 15th Jun 2023, so that I can compare the before and after.

3.2. Discussion Volume

First, to understand if the new Abarth 500e is bringing up more discussion on Reddit, the number of texts is plotted against time.

First, the monthly count is plotted against time. Even though the earlier data dates back to June 2013, the discussion volume before 2019 is not significant, so only volumes from 2019 onwards is plotted.

[7] https://www.media.stellantis.com/uk-en/abarth/press/abarth-500e-to-make-uk-public-debut-at-salon-prive-london

As we can see in *Figure 3*, there is not any clear association between discussion volume and the Announcement & Showcase. Even though there is a peak in volume in the Release month (2023 Jun), we cannot conclude the Release led to more discussion, as causation requires much more evidence.

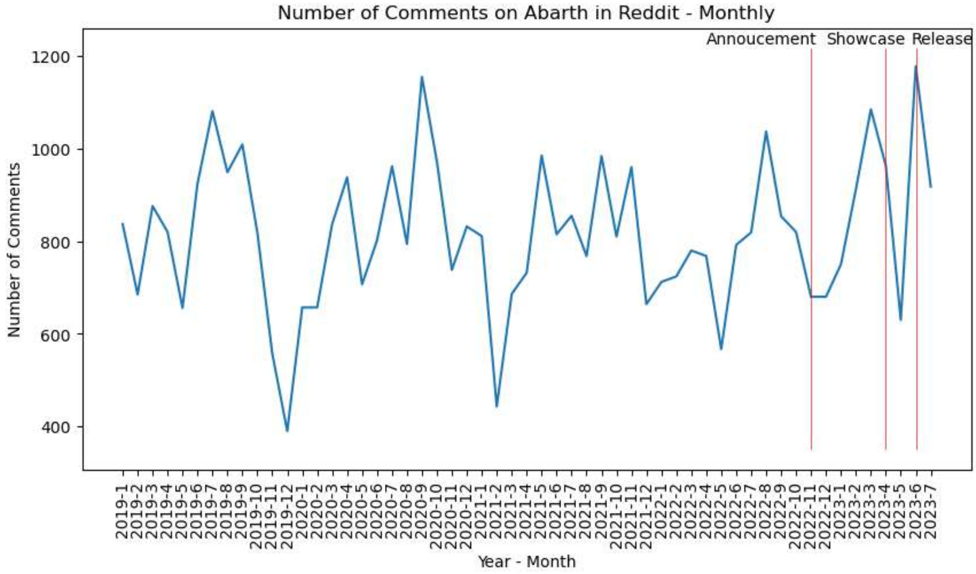

Figure 3: Abarth Discussion Volume against Year-Month from 2019 onwards

Deep diving into weekly volumes around the significant dates in *Figure 4*, we can see peaks around the Showcase week and Release week. Still, we cannot conclude the two events led to more discussion; it is possible that the Abarth marketing team created more posts to raise attention.

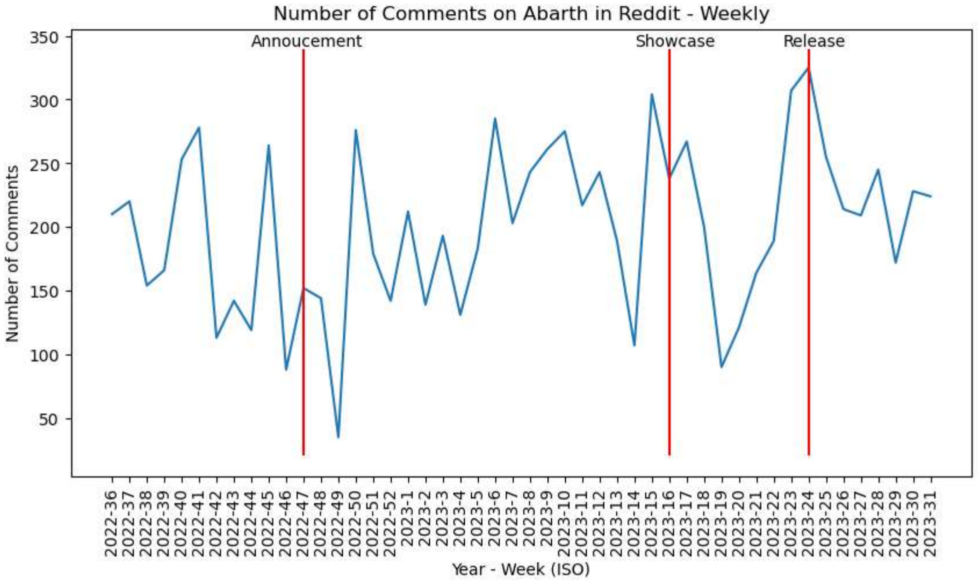

Figure 4: Abarth Discussion Volume against ISO Year-Week from 2022 September onwards

3.3. Meaning System

In understanding the meaning system attached to Abarth, we extracted the main discussion topics through Topic Modelling, and looked at the keywords in each topic. First, we have to decide the optimal number of topics and the best topic model. Then, we can generalize the topics and give them a label. Finally, we can analyse the topics, keywords, and the distribution over time, especially before and after the significant dates.

3.3.1. Topic Modelling

After cleaning, preprocessing and basic filtering of the dataset, the ten most frequent words were filtered out. I then created multiple topic models and selected the optimal one based on the coherence score (interpretability) and topic quality (if the topics are distinct enough).

Model 1

Based on the Coherence Score plot (*Figure 5*), the model with five topics is the most interpretable, having the highest score.

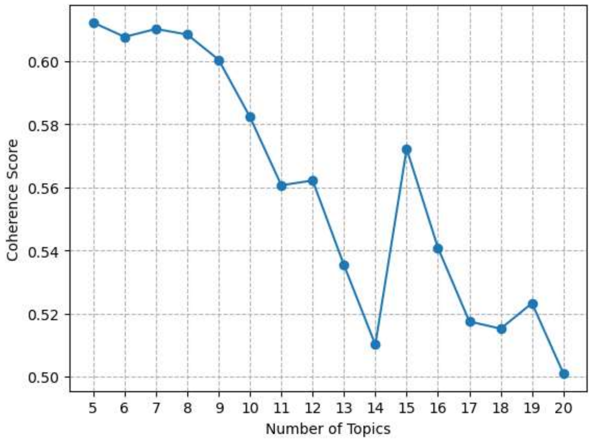

Figure 5: Coherence Score VS Number of Topics plot after filtering out 10 most frequent words

The Intertopic Distance Map (*Figure 6*) shows that topic 0 and 2 are extremely similar while have three overlapping words: 'want', 'good', and 'know' (*Table 1*). Those words are among the most frequent words in the whole dataset, alongside with the ten I had already filtered out earlier. They do not help in understanding the topics.

From this, I decided to filter the three words out from the whole dataset for modelling.

Topic	Most Frequent Words							
0	want	good	wheel	love	well	tire	new	know
1	work	time	light	start	issue	try	turn	battery
2	year	good	find	know	want	mile	model	lot
3	engine	oil	need	use	sure	issue	turbo	power
4	replace	mile	warranty	issue	new	dealer	need	repair

Table 1: Most Frequent Words in the 5-topics topic model

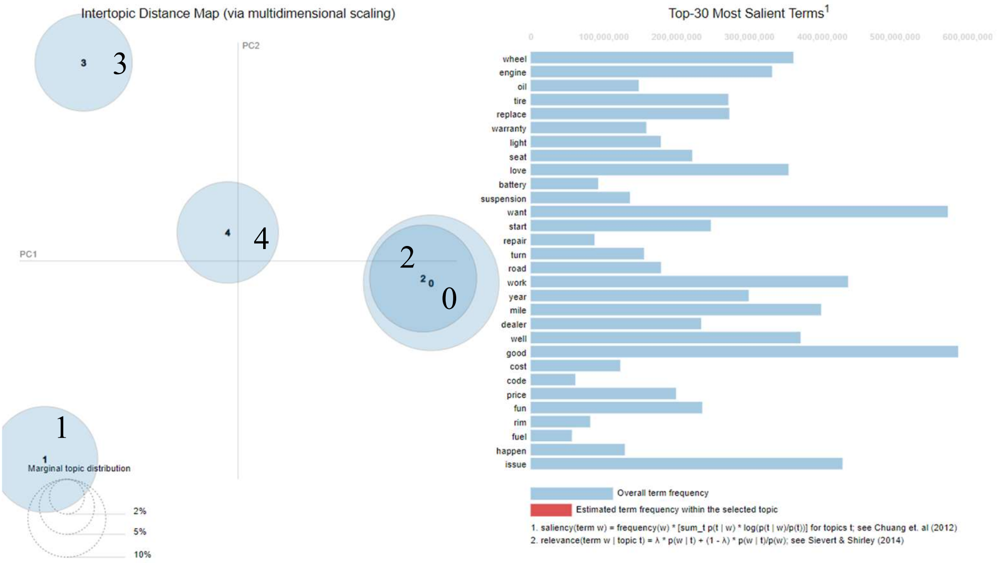

Figure 6: pyLDAvis visualisation for the 5-topics model (10 words filtered out)

Model 2

After having the dataset filtered three extra words out, again, multiple models were built. According to the new Coherence Score plot (*Figure 7*), 8-topics model is the most interpretable one. However, having eight topics is a bit too much: each topic is not distinct enough, particularly for topic 2 and 5, and topic 0, 1 and 6, as we can see in *Figure 8*.

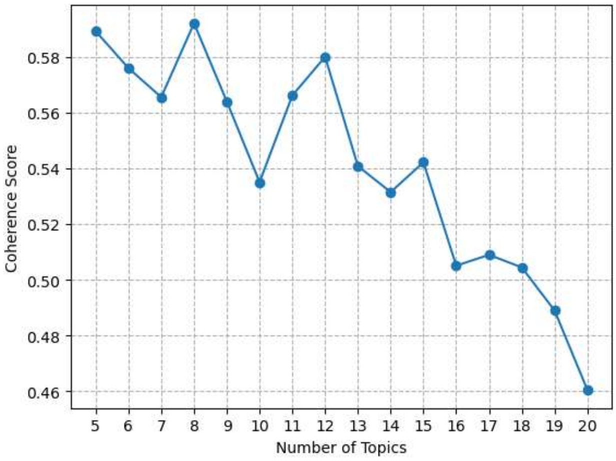

Figure 7: Coherence Score VS Number of Topics plot after filtering out 10 most frequent words + 3 extra words

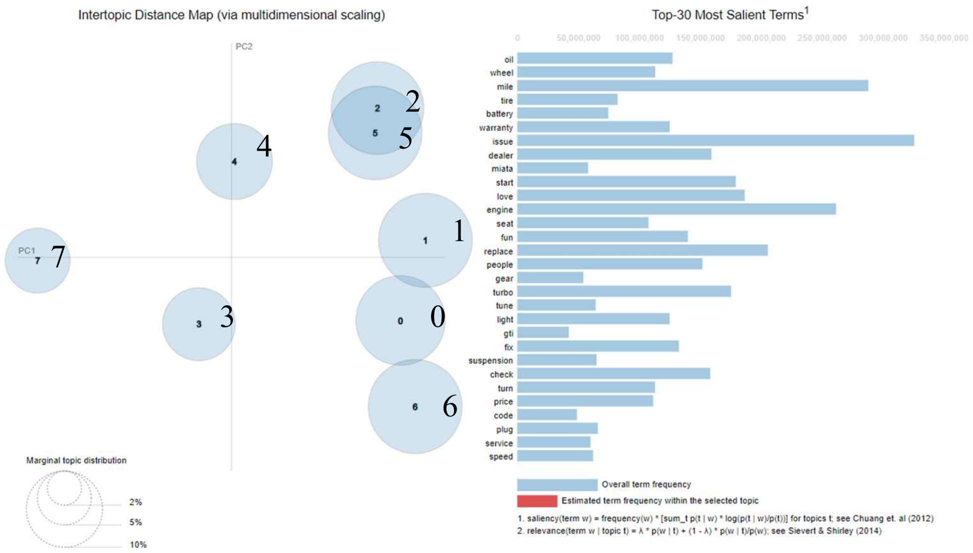

Figure 8: pyLDAvis visualisation for the 8-topics model (13 words filtered out)

Model 3

Then, I assessed the next best model, which is the 5-topics model (*Figure 7*). Having only five topics, two groups of two topics are very close together in the two-dimensional space (*Figure 9*). It would not be detailed enough in understanding the meaning system with only three distinct topics (i.e., 0&1, 2&4, 3).

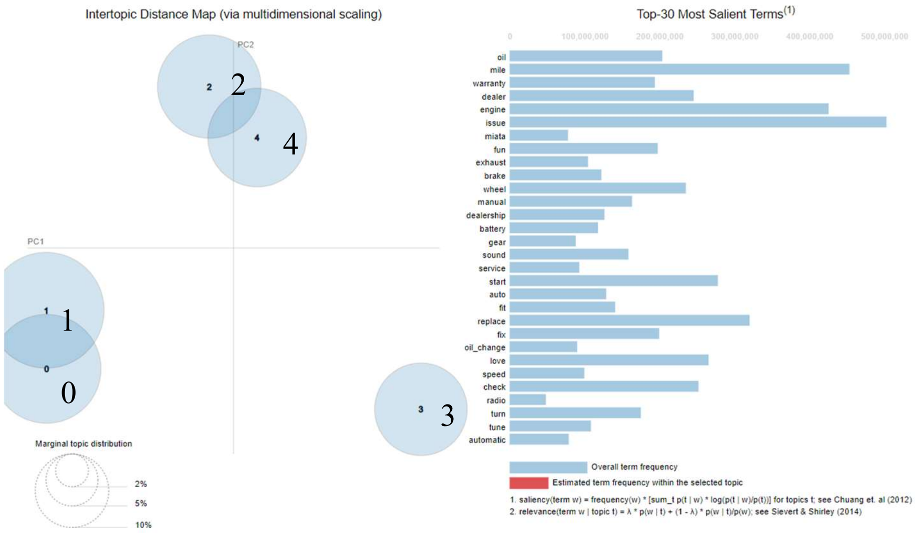

Figure 9: pyLDAvis visualisation for the 5-topics model (13 words filtered out)

Final Model

With the second-best model being not good enough, I assessed yet another model. The 12-topics model is the third best (*Figure 7*), however, having eight topics is already too much, not to mention twelve. The next best model while having less topics is 6-topics one.

The topics are generally quite distinct, only topic 0 and 1 are similar (*Figure 10*). The topics are distinguishable by the frequent words (*Table 2*) in some sense.

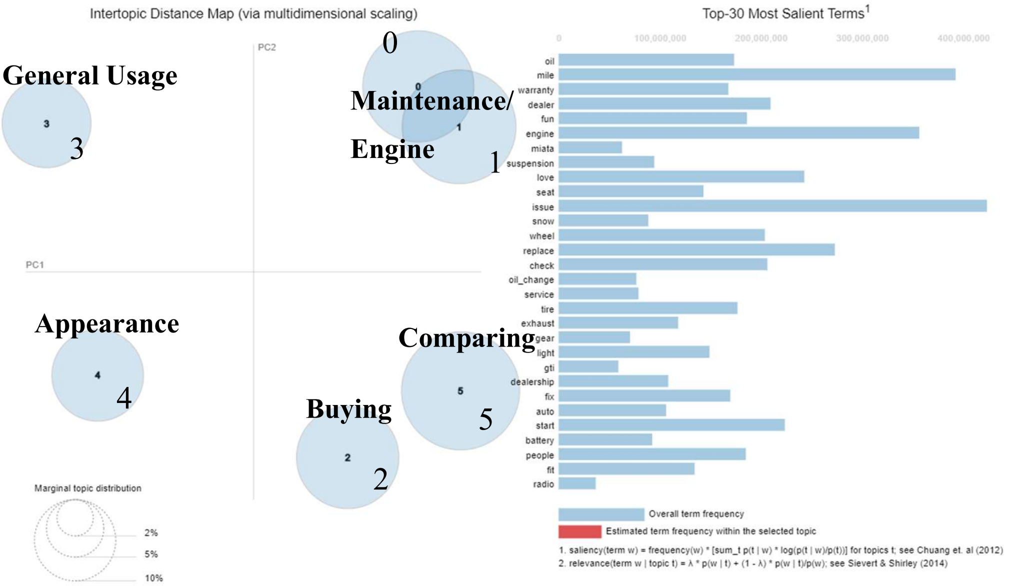

Figure 10: pyLDAvis visualisation for the 6-topics model (13 words filtered out)

Topic	Most Frequent Words							
0	mile	oil	dealer	issue	replace	engine	warranty	check
1	time	issue	engine	start	turn	light	run	work
2	people	find	work	love	miata	price	time	great
3	work	tune	radio	sure	model	use	need	new
4	wheel	find	model	light	red	black	color	one
5	new	love	well	fun	turbo	mile	need	tire

Table 2: Most Frequent Words in the 5-topics topic model (13 words filtered out)

To fully understand the topics, I examined some examples from each topic: each having a high doc-to-topic probability. The topics can be concluded as the labels shown in *Figure 10*. Topics 0 and 1 were combined into one, as the examples agreed with their closeness in the Intertopic Distance Plot (*Figure 10*).

3.3.2. Topic Distribution

With the defined topics, the topics volumes were plotted against time to understand the discussion focus shift along with the significant dates.

Overall, the most discussed topic was 'Maintenance/Engine', as shown in *Figure 11*, making up around 30% of the discussion. The least discussed topic was 'General Usage', making up only 10%.

Pre- and post- significant dates, there is no visible change in topic distribution, neither in monthly level (*Figure 11*) nor in weekly level (*Figure 12*). The boxplot of weekly topic proportion (*Figure 13*) proves the absence of significant change. It means that people's focus is constant along time.

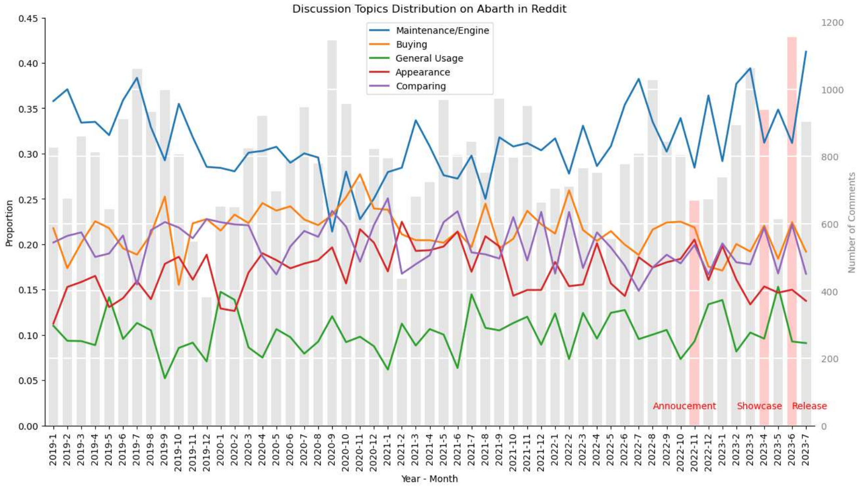

Figure 11: Abarth Discussion Topic Distribution against Year-Month from 2019 onwards

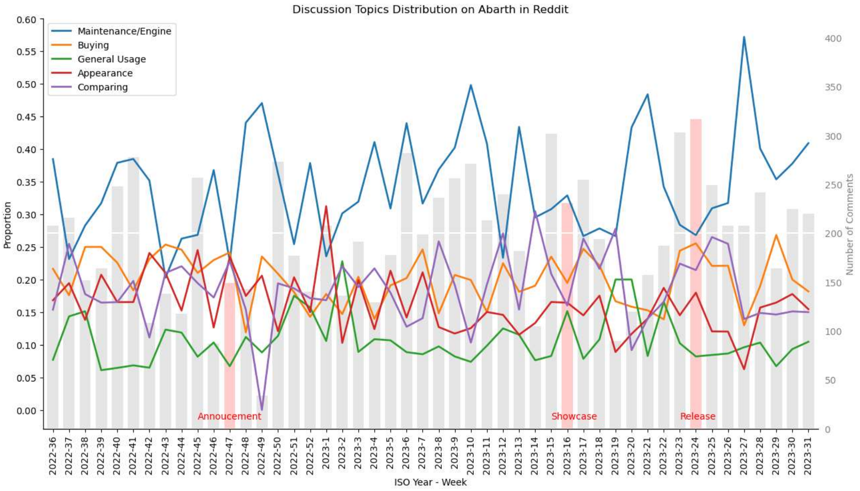

Figure 12: Abarth Discussion Topic Distribution against ISO Year-Week from 2022 week 36 onwards

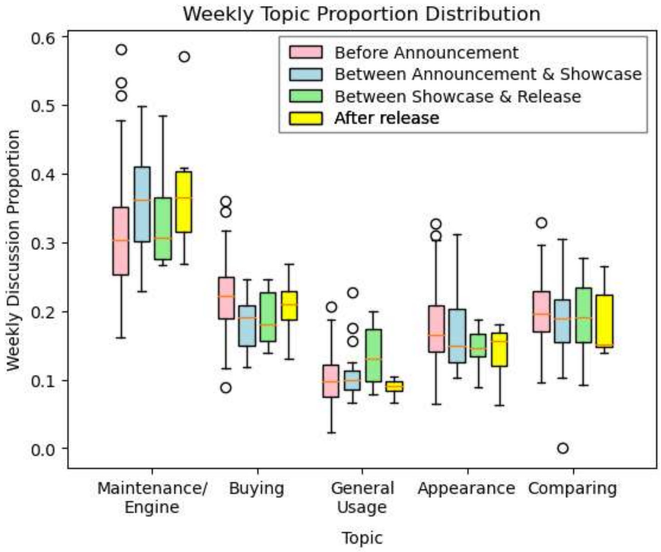

Figure 13: Boxplot of Topic Proportion grouped by Pre- and Post- Significant Dates

3.3.3. Topic Keywords

To understand each discussion topic more in-depth, the evolution of the ten most frequent words is plotted for each topic, to see if the focus shifted within a topic. Only top ten for before Announcement and after Release were plotted to avoid confusion.

Among the five topics, only Maintenance/Engine was relatively stable in terms of keywords and their rankings (*Figure 14*). Nine of the top ten words before Announcement remained in the top ten after Release. Also, the rankings of the words did not change much. While for other topics, especially for General Usage (*Figure 14c*), the keywords before Announcement and after Release were so different; the rankings were also more volatile.

From these changes, particularly for General Usage, we can see that people were focusing on the sound of the car engine more, with the words ‘sound’, ‘exhaust’ and ‘valve’ dominating at the end. This makes so much sense as one of the key features of the new Abarth 500e is recreating the iconic engine sound.

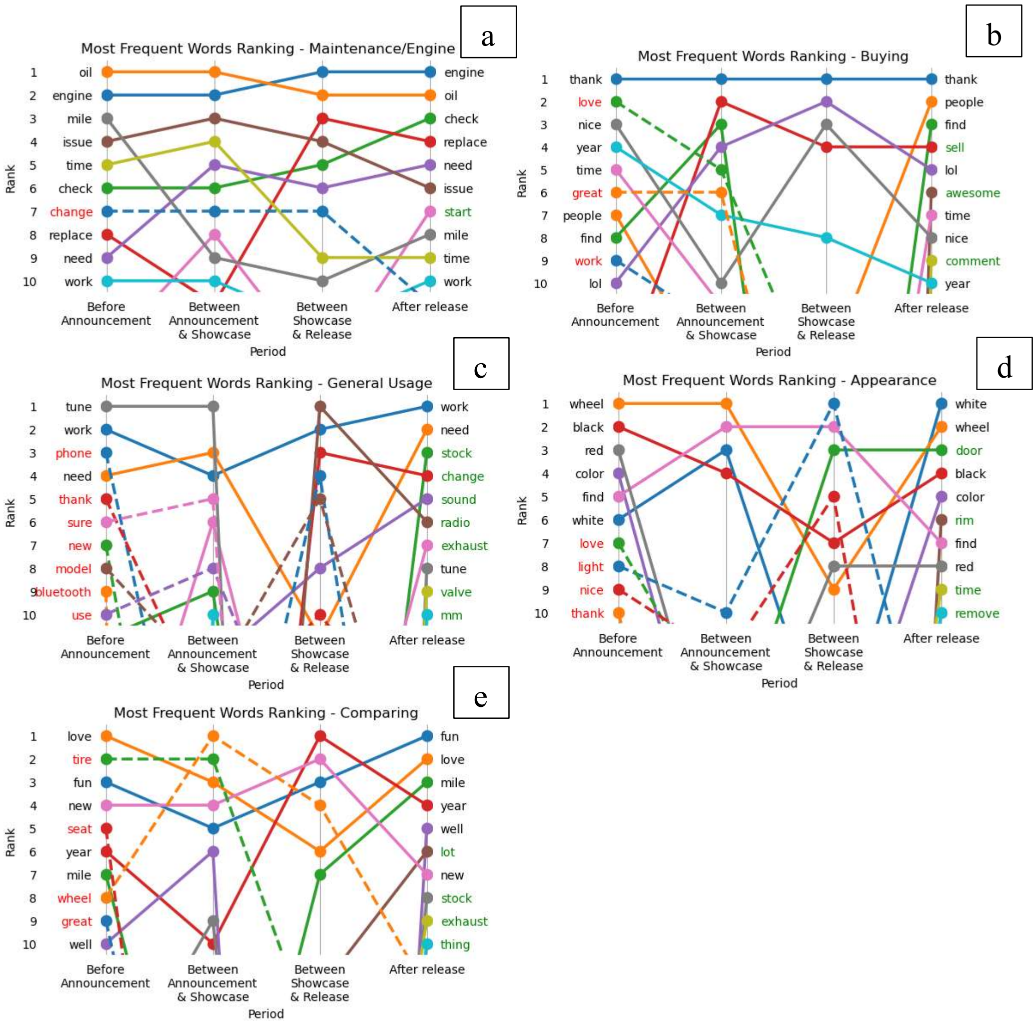

Figure 14: The 10 Most Frequent Words Ranking by Topic

3.4. Discussion Sentiment

With the topics generated, we can understand people's emotions towards each of them with Sentiment Analysis. To understand the general sentiments and the changes after the significant dates, the weekly sentiment proportion distributions were plotted by topic (*Figure 15*).

Overall, we can see that for three topics: Maintenance/Engine, General Usage, and Appearance (*Figure 15a, c, d*), people were neutral in general (with the median proportion of neutral greater than 75-percentile of positive/negative). For Comparing (*Figure 15e*), people were quite positive, having similar weekly proportion for neutral and positive. While for Buying (*Figure 15b*), the discussions were in general positive.

Deep diving into each topic, overall, the sentiments did not change much before and after the three significant dates. However, there are two points worth noticing:

1) for Maintenance/Engine, less discussions were positive after the Announcement (under 'positive', all three median proportion after Announcement were below the 25-percentile before Announcement (red line)) (*Figure 15a*);
2) for Maintenance/Engine, there were more negative discussions than positive ones in all periods (Figure 14a);
3) for General Usage, more discussions were negative after the Release (under 'negative', the median proportion after Release was above the three 75-percentiles before Release (red line)) (*Figure 15c*).

Referring back to Section 3.3.3 (Page 23), General Usage was more about the car engine sound. With point 1 and 3 above, we can say that people in Reddit discussing Abarth were not so positive about the engine sound of the new Abarth 500e. Even though Abarth did address this point by making an artificial sound generator (Pappas, 2023), I think the audience are still doubtful about the absence of a real engine and its iconic sound.

Furthermore, with point 2, we can say that people have not been satisfactory on the Maintenance/Engine.

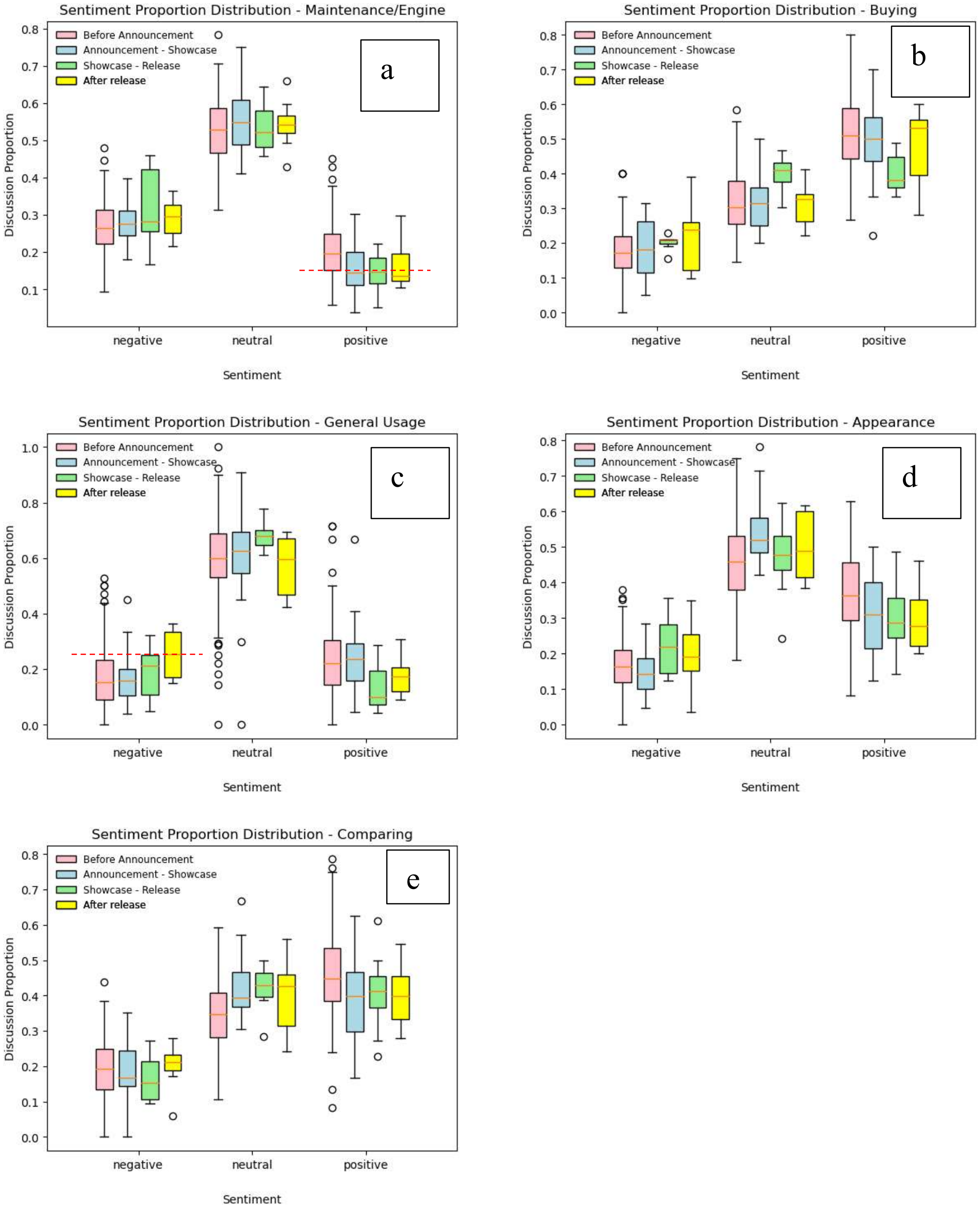

Figure 15: Boxplot of Sentiment Proportion grouped by Pre- and Post- Significant Dates by Topic

4. Conclusion & Discussion

Topic Modelling and Sentiment Analysis were used to understand the online discussion on Abarth. In Reddit, people usually talked about five topics: 1) Maintenance/Engine, 2) General Usage, 3) Appearance, 4) Buying, and 5) Comparing. Each topics' focus did not change much across time, before and after the EV launch, except for General Usage: it focused on the engine sound more after the Market Release of the new Abarth 500e.

The sentiment for each topic was generally stable across time. For Maintenance/Engine, discussions were mainly neutral, with more negative than positive. For General Usage, discussions were mainly neutral, with similar proportion for negative and positive. For Appearance, discussions were also mainly neutral, with more positive than negative. For Buying, discussions were positive in general. For Comparing, discussions were mostly neutral and positive.

There are three main insights from the topic keywords and by-topic sentiments:

1) The Reddit discussions are consistent with Abarth's sought identity. From the topic keywords, particularly for Comparing (*Figure 14e*), we can see words 'fun' and 'love'. This is directly addressing Abarth's top identity: 'fun'.
2) Maintenance/Engine and General Usage discussions were less positive and more negative after the electrification.
3) Even though Maintenance/Engine discussions were neutral in general, there were more negative discussions than positive ones.

From the latter two points, there arises two possible challenges to Abarth:

1) People are doubtful towards the new Abarth 500e on the absence of a real engine and its iconic sound. Abarth's legacy would be threatened, and people would be less willing to buy the new EV.

2) People were in general not satisfied with the general performance or maintenance of traditional combustion engine Abarth. If Abarth is trying to make people accept their new EV, which does not have the iconic engine and its sound, with a high performance, it would be a difficult task.

Below are the recommendations to address the two possible challenges respectively:

1) Perfecting the artificial sound generator, which Abarth is doing now (Pappas, 2023), and developing a vibration generator to mimic the engine vibration.

   These two ways can help to give the Abarth users an experience as if the engine is still here.

2a) Improving the quality and performance of the battery and making maintenance more accessible.

   This directly addresses the unsatisfaction on Maintenance/Engine: the direct equivalent of engine for EV is battery. A higher quality and performance vehicle brings less maintenance issue. A more accessible maintenance method, together with lower maintenance effort, would drastically improve user experience.

2b) Improving the design of the interior hardware and software.

   This directly addresses the unsatisfaction on General Usage, as words like ‘radio’, ‘phone’, ‘system’, and ‘bluetooth’ appeared very frequently under this topic (*Figure 14c*). It helps to improve the overall user experience, making people more satisfies with the new Abarth 500e.

Appendix

Variable	Description
title	The title of the Reddit post
author	The username of the author of the Reddit post
text	The textual content of the Reddit post
numcm	The number of comments under the Reddit post
ups	The number of up votes for the Reddit post
downs	The number of down votes for the Reddit post
score	The number of up votes net of down votes for the Reddit post
link	The URL link of the Reddit post
time	The creation time of the Reddit post in Unix time format
id	The exclusive identifier of the Reddit post

Table 3: Data Dictionary of Post Data Scraped with Reddit API

Variable	Description
title	The title of the Reddit post
author	The username of the author of the Reddit post
text	The textual content of the Reddit post
numcm	The number of comments under the Reddit post
score	The number of up votes net of down votes for the Reddit post
up_ratio	The ratio of up votes to down votes for the Reddit post
link	The URL link of the Reddit post
time	The creation time of the Reddit post in Unix time format
id	The exclusive identifier of the Reddit post

Table 4: Data Dictionary of Post Data Scraped with Pushshift API

Variable	Description
author	The username of the author of the Reddit comment
text	The textual content of the Reddit comment
score	The number of up votes net of down votes for the Reddit comment
time	The creation time of the Reddit comment in Unix time format
id	The exclusive identifier of the Reddit comment
parent id	The identifier of the comment showing the hierarchy and the id of its parent Reddit post/comment (eg t1_xxx: the comment is replying to the comment of id xxx; t3_xxx: the comment is replying to the post of id xxx)
post id	The exclusive identifier of the parent Reddit post of the comment

Table 5: Data Dictionary of Comment Data Scraped with Reddit API

References

[1] Bansal, S. (2016) *Beginners Guide to Topic Modeling in Python*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/#h-endnotes (Accessed: 20 August 2023).

[2] Consumer Reports (2013) *Tesla Model S Review: An electric sports car earns our top test score*, *Consumer Reports*. Available at: https://www.consumerreports.org/cro/magazine/2013/07/tesla-model-s-review/index.htm (Accessed: 24 August 2023).

[3] International Energy Agency (2023) *Global EV Outlook 2023 - Executive summary*, *International Energy Agency*. Available at: https://www.iea.org/reports/global-ev-outlook-2023/executive-summary (Accessed: 24 August 2023).

[4] Laukkonen, J. (2022) *What Is a Subreddit?*, *Lifewire*. Available at: https://www.lifewire.com/what-is-a-subreddit-5271885 (Accessed: 19 August 2023).

[5] Pappas, T. (2023) *Abarth Fakes It 'Til They Make It: The 500e's New Artificial Sound Generator*, *Carscoops*. Available at: https://www.carscoops.com/2023/04/it-took-abarth-two-years-to-come-up-with-the-fake-sounds-of-the-500e/ (Accessed: 28 August 2023).

[6] Patton, M. (2022) *Competition Heats Up For Tesla In EV Market*, *Forbes*. Available at: https://www.forbes.com/sites/mikepatton/2022/03/30/competition-heats-up-for-tesla-in-ev-market/ (Accessed: 24 August 2023).

[7] Philander, K. and Zhong, Y. (2016) 'Twitter sentiment analysis: Capturing sentiment from integrated resort tweets', *International Journal of Hospitality Management*, 55, pp. 16–24. Available at: https://doi.org/10.1016/j.ijhm.2016.02.001.

[8] Rapier, R. (2022) *Why Tesla's Market Share Is Set To Plunge In 2023*, *Forbes*. Available at: https://www.forbes.com/sites/rrapier/2022/12/19/why-teslas-market-share-could-plunge-in-2023/ (Accessed: 19 August 2023).

[9] Standage, T. (2021) 'The lost history of the electric car – and what it tells us about the future of transport', *The Guardian*, 3 August. Available at: https://www.theguardian.com/technology/2021/aug/03/lost-history-electric-car-future-transport (Accessed: 24 August 2023).

[10] Stellantis (2023) The 'roar' of the New Abarth 500e designed by the Stellantis Sound Design Studio: more than 6,000 hours to recreate the perfect sound, Stellantis. Available at: https://www.media.stellantis.com/em-en/abarth/press/the-roar-of-the-new-abarth-500e-designed-by-the-stellantis-sound-design-studio-more-than-6-000-hours-to-recreate-the-perfect-sound (Accessed: 24 August 2023).

[11] Stokes, D.C. *et al.* (2020) 'Public Priorities and Concerns Regarding COVID-19 in an Online Discussion Forum: Longitudinal Topic Modeling', *Journal of General Internal Medicine*, 35(7), pp. 2244–2247. Available at: https://doi.org/10.1007/s11606-020-05889-w.

[12] Teja, S. (2020) *Stop Words in NLP*, *Medium*. Available at: https://medium.com/@saitejaponugoti/stop-words-in-nlp-5b248dadad47 (Accessed: 19 August 2023).

[13] Zvornicanin, E. (2021) *When Coherence Score is Good or Bad in Topic Modeling?*, *Baeldung*. Available at: https://www.baeldung.com/cs/topic-modeling-coherence-score (Accessed: 20 August 2023).