Preliminary discovery report

Table of Contents

Objective 1

Note 1

I. Context 2

II. Data methodology 2

Data collection 2

Data processing 3

Descriptive analytics 3

III. Analyses 5

Dynamic sentiment analysis 5

Topic modelling 7

Document similarity 8

IV. Deep dive into select periods and news 8

V. Building a predictive model 10

Connecting new articles to sentiment at that time 10

Building our model 11

Limitations of our model 11

VI. Testing our model vs recent news 12

Cross-validation 12

Real-world example 12

VII. Key commercial implications and strategies 12

VIII. Conclusion remarks 14

Limitations of our research 14

Discussion 14

X. References 15

Scope

We seek to analyze autonomous vehicles sentiment from Reddit users across time, extract key topics and attributes, and predict sentiment based on recent news' labeled topics. We seek to recommend design, engineering, marketing and PR strategies for car manufacturers and autonomous vehicle software companies to facilitate better end-user adoption.

I. Context

Projected to generate between $300B and $400B in revenues by 2035 (McKinsey, 2023), autonomous vehicles (AVs) have attracted considerable attention from both tech optimists and pessimists conflicted in their visions of the future, but also city officials and urban citizens looking for more efficient mobility solutions. With Big Tech, incumbent and new automakers affirming their commitments to commercialize self-driving cars through large R&D investments, and partnership announcements and press releases, AVs' wide adoption and success ultimately rests on end users of this technology, the consumers.

With the rapid spread and democratization of information and communication technologies, the world is more connected than ever, generating a wealth of data for companies to analyze customer sentiment and opinions, notably on social media platforms. Given the dependence AV makers have on end-user adoption and their perception of these cars, companies can find a great interest in both understanding user sentiment on their brands but also assessing this sentiment across times as news and key events relating to progress or failures happen in their development.

As more solutions are tested in cities across the U.S. (San Francisco and Phoenix with Uber, Tesla, Cruise and Waymo notably) and China (Beijing with Baidu), consumers remain observant of company missteps, informing on key product and software designs to ease commercialization and gain an edge once the technology is nationally validated by regulators and ready for mass-scale adoption.

Previous bodies of research have undergone the task of analyzing user sentiment and perception of driverless cars and technologies to replace traditional mobility. Gupta and Sharma (2022) used text mining procedures to understand better user sentiment on AVs, concluding that the public was neither positive nor negative on the technology, thus informing companies that more proof of concepts and use cases were necessary to build excitement and competitive edge. This is further reinforced by Ding, Korolov, Wallace, and Wang (2021) who, by also leveraging Twitter feeds, noted potential policy designs but also the inherent sensitivity of user sentiment to certain news and events, highlighting the role the media has in AV acceptance.

Our study complements previous research on social media platform user sentiment on AVs by (a) introducing a dynamic component by observing sentiment across time and key news events, (b) correlating key user-discussed topics with shifts in user sentiment, and (c) predicting user sentiment based on future news' underlying topic assessment.

Our purpose is to support corporate decisions on expenditure and support engineering, product design and marketing strategies for commercial success. Additionally, we aim to build a tool for company marketers to test the sensitivity of user sensitivity to specific events and news across key topics and dimensions.

II. Data methodology

Data collection

While there are a plethora of social media platforms to choose to scrape user opinion from, we selected Reddit because of Twitter's recent ban on free web scraping (the API price being too

considerable for us to bear), its superiority as a post-based and opinion platform with a clear data architecture facilitating both extraction and understanding and its deep size with thousands of subreddits and discussion channels enabling the extraction of hundreds of thousands of user comments. This rendered Reddit the best and most simple solution to use. To scrape the data, we used the PRAW package and wrote scripts enabling us to scrape the top 1,000 posts and their associate comments across 4 subreddit channels of our choosing: r/SelfDrivingCars (by far the largest community discussing the topic), r/Waymo, r/AutonomousVehicles, and r/SelfDrivingCarsLie. Combining data from 4 different channels was designed to counter the query limit of 1,000 posts set by Reddit, allowing us to aggregate more and older data points[1]. Our scraping code can be found in the *Scripts* folder in the *Data.zip* submitted.

While there are a plethora of social media platforms to choose to scrape user opinion from, we selected Reddit because of Twitter's recent ban on free web scraping (the API price being too considerable for us to bear), its superiority as a post-based and opinion platform with a clear data architecture facilitating both extraction and understanding and its deep size with thousands of subreddits and discussion channels enabling the extraction of hundreds of thousands of user comments. This rendered Reddit the best and most simple solution to use. To scrape the data, we used the PRAW package and wrote scripts enabling us to scrape the top 1,000 posts and their associate comments across 4 subreddit channels of our choosing: r/SelfDrivingCars (by far the largest community discussing the topic), r/Waymo, r/AutonomousVehicles, and r/SelfDrivingCarsLie. Combining data from 4 different channels was designed to counter the query limit of 1,000 posts set by Reddit, allowing us to aggregate more and older data points[2]. Our scraping code can be found in the Scripts folder in the Data.zip submitted.

Data processing

Having scraped comments, we went through the following Excel processing steps to derive a useable dataset for analysis:

1. Removing blank comment text cells
2. Removing duplicate entries
3. Correcting Excel column datatypes (ensuring dates are treated as such)

After completing these steps, we were left with 23,965 comments making up our final dataset as shown in the '*Final master*' tab.

Descriptive analytics

After processing, we proceeded to plot our dataset across select dimensions to understand the underlining data structure. First, we evaluated the distribution of comments across the years through a pivot count as illustrated in Exhibit 1. While we can see that 82.85% of our dataset rests on comments from 2023, we observe that 2021 was an active year for user discussions, potentially hinting at major developments.

---

[1] r/SelfDrivingCars is so frequent that scraping 1,000 posts ends up with data over the last 6 months compared to other channels who go back all the way until 2017 in certain occasions.

[2] r/SelfDrivingCars is so frequent that scraping 1,000 posts ends up with data over the last 6 months compared to other channels who go back all the way until 2017 in certain occasions.

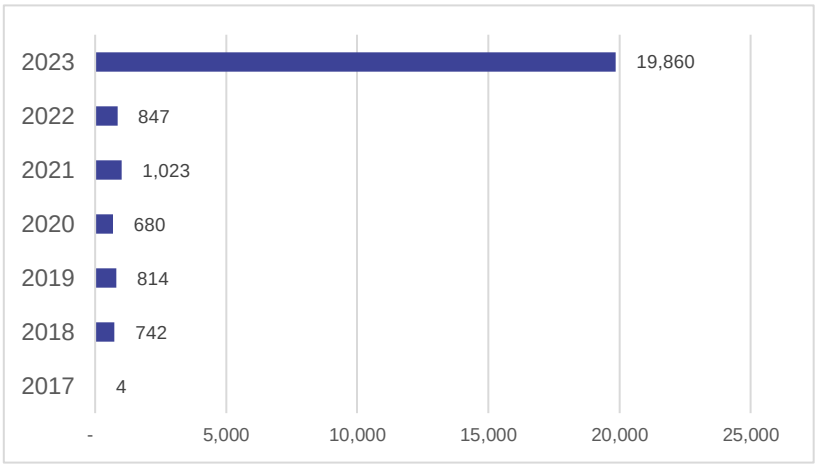

Exhibit 1 – Yearly distribution of comments

Second, we looked at the distribution of all comments across the four subreddit channels, confirming our concerns that r/SelfDrivingCars concentrates most (79.64%) of our dataset. This might raise dimension issues, which we neglect given that subreddit belonging is trivial in our study. Exhibit 2 illustrates the distribution.

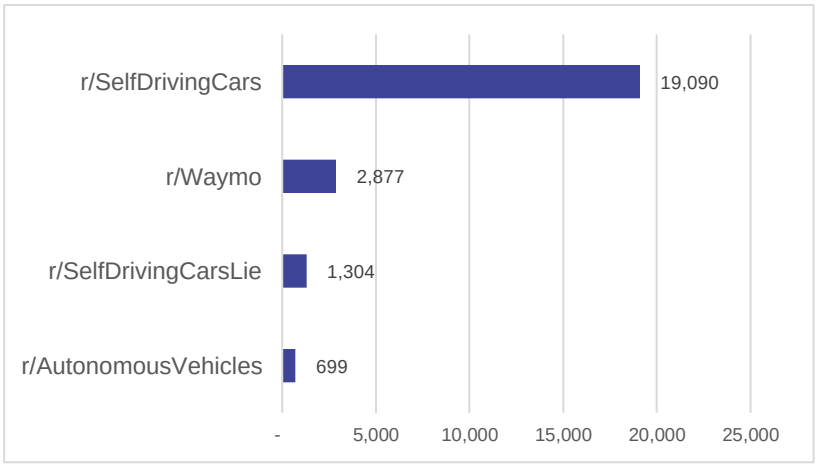

Exhibit 2 – Comment distribution by subreddit channel

Finally, we incorporated both views by plotting the monthly distribution of comments by subreddit channel to identify where and when each comment was generated. Unsurprisingly, most 2023 comments were concentrated in r/SelfDrivingCars which legitimized why pulling data from multiple subreddits was important to ensure representation through time. Exhibit 3 displays different views of this distribution (a filter in the '*Monthly distribution*' tab can be used to replicate).

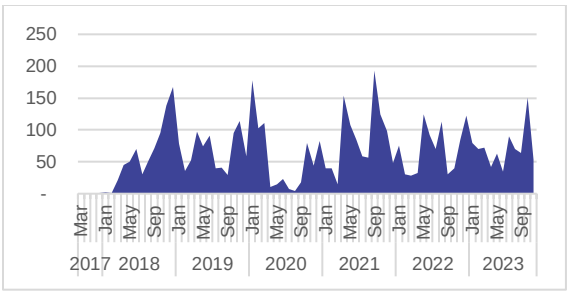

Exhibit 3a – Monthly distribution across subreddit channels, without r/SelfDrivingCars

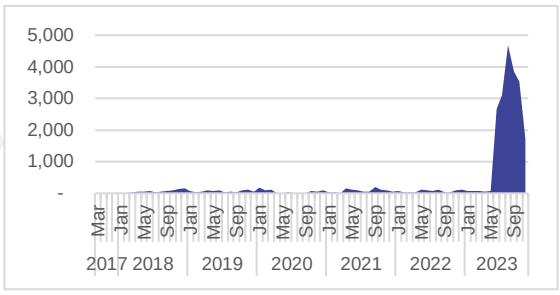

Exhibit 3b – Monthly distribution across subreddit channels, including r/SelfDrivingCars (undifferentiated)

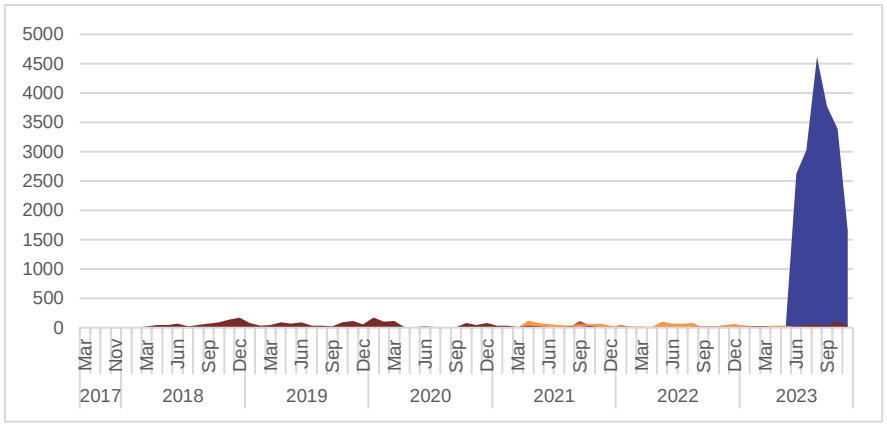

Exhibit 3c – Monthly distribution across subreddit channels

III. Analyses

Dynamic sentiment analysis

Our initial goal was to analyze user sentiment on the topic of autonomous vehicles across time. As our data spans from March 17th, 2017, to November 12th, 2023, we hypothesized that there would be a moderate sentiment shift across the period.

First, we processed once more the dataset by removing irrelevant columns that would not assist our analysis[3]; removed cells with blanks, '[deleted]', and '[removed]'; and replaced ">" with " " as the sign was initially meant to show a comment reply (useless for our analysis).

Second, with the cleaned dataset, we ran TextAnalyzer's NRC Hashtag Emotion and Sentiment analysis with all the comments and set a threshold that comments within [-0.1;0.1] would be replaced to 0 ('updated_SENT' column), representing a comment of neutral sentiment. A sample output is found in Exhibit 4.

[3] We only maintained columns with the comment text and date

CLEANED_comment_text	SENT_comment_date	SENT	updated_SENT
"Thanks, it's delicious."	2023-06-09	0.5395	0.5395
Aren't you the same guy who mischaract	2023-06-09	-0.12214	-0.12214
We have evidence from the owner of the	2023-06-09	-0.07312	0
Gosh. Tesla Stans responding to every t	2023-06-10	-0.22025	-0.22025
This is relevant to Tesla conduct how?	2023-06-09	-0.36314	-0.36314
1. This incident wouldn't be reported per	2023-06-09	-0.19987	-0.19987
I don't think hitting a pet is the purpose of	2023-06-09	-0.18101	-0.18101
Correction: We have evidence from the c	2023-06-09	-0.05023	0
Not sure why you're putting Beta in quote	2023-06-09	-0.12732	-0.12732
No one says ro move over. How about s	2023-06-09	-0.20246	-0.20246
People on this sub say a failure to predict	2023-06-09	-0.097	0
They said the dog came from behind a p	2023-06-09	-0.12465	-0.12465
FSD "Beta" being released to consumers	2023-06-09	-0.1781	-0.1781
Tesla didn't release this. A user did, other	2023-06-09	-0.17072	-0.17072
Their CEO lies and lies. So who's fault re	2023-06-09	-0.22824	-0.22824

Exhibit 4 – TextAnalyzer sample output

Third, we plotted comment updated sentiment scores on a weekly basis as we noticed that this basis provided the most insightful and granular representation of sentiment shifts. Exhibit 5 presents the resulting weekly sentiment distribution[4] including a trendline with a slope of a = -0.00005 which, although positive given the inversion, is small enough to be considered negligible.

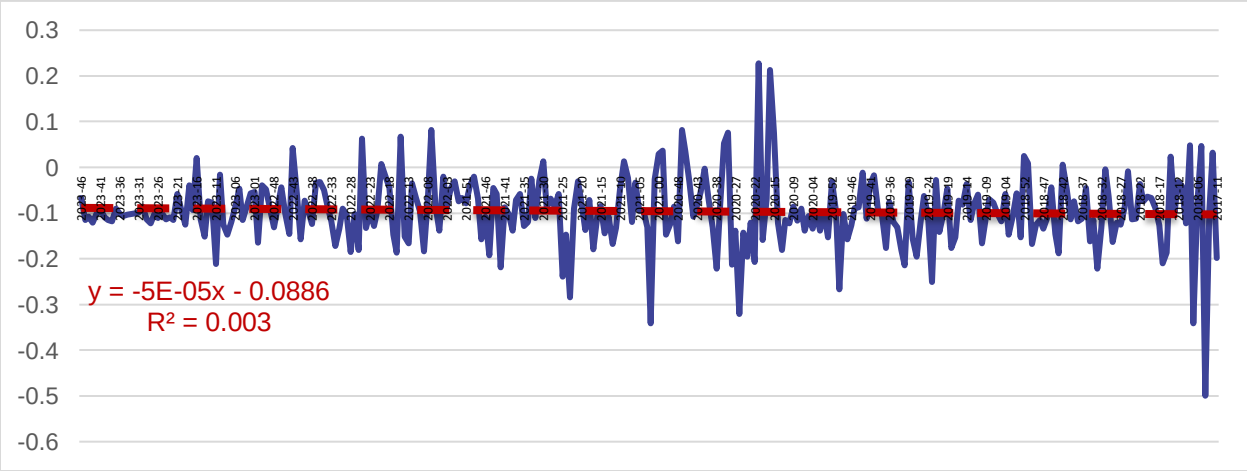

Exhibit 5 – Weekly updated sentiment distribution

Exhibit 5 provides 3 key insights:

1. Between 2017 and 2023, sentiment on autonomous vehicles does not seem to have progressed meaningfully when analyzed at the macro level. Users appear to be just as likely to like AVs in both periods. With an average sentiment score of -0.0958, it appears that Reddit users are on average critical about AVs, but their opinion is slowly turning positive as informed by the low trendline. This is consistent with a thesis that users were skeptical about AVs and require strong conviction before mass adoption is needed.
2. User sentiment is significantly more stable in the last 6 months of 2023 compared to a more volatile and noisier period before June 2023 (around the 21st-26th week of 2023).

[4] The x-axis is inverted: points closer to zero are more recent.

This can be explained by the concentration of our data points in that period highlighting that the volume of data enables us to draw a more consistent picture.

3. User sentiment appears to have shifted tremendously around specific periods which we rationalize as times bearing critical news (both positive and negative). This provides deep-dive cases for further analysis by scraping news data for these specific months or weeks. We have identified the following for further analysis:
   a. December 2018
   b. January 2020
   c. March 2020
   d. January 2021

Given those insights, we now split our research into different streams by investigating (i) daily user sentiment in the last 6 months for more variability to trace back to specific news on high sentiment shift days/periods, and (ii) news which happened during each of the four aforementioned periods. Section IV provides an overview of the identified news in these periods.

Topic modelling

In parallel, we conducted topic modelling on our comment text to extract and understand better key topics that users are discussing on AVs. After running LDA with a different number of topics each time, we settled on a total number of 9 topics. Our process to evaluate if topics made sense was to evaluate the top 25 words per topic for each iteration and see if the group was homogenous. Similar to Assignment 2, we used ChatGPT to give us topic labels, which we then assessed with human validation. Exhibit 6 illustrates the 9 topics and their respective top 25 words. These topics were: 'FSD' (Tesla's Full Self-Driving), Assistance and Traffic Control Systems (in essence, discussions on traffic), 'Public Needs' (what the general public seeks and demands), 'Foreign players' (mentions of non-Big Tech players, notably overseas players), 'Safety', 'Big Tech players', 'Transit and commute', 'Sensor technology' (including LIDAR), and 'City trials'.

Word	FSD
tesla	0.04015838
driving	0.02704766
fsd	0.01911994
driver	0.01672163
self	0.0155891
car	0.01524268
level	0.0116319
system	0.01155196
even	0.0082876
safety	0.00779462
drive	0.00771467
cars	0.00715507
autopilot	0.00648888
human	0.00627569
like	0.00539631
think	0.00530305
drivers	0.00524975
still	0.00516981
take	0.00502325
2	0.00487668
time	0.00482339
beta	0.00482339
wheel	0.00465018
autonomous	0.00458356
would	0.00447697

Word	Assistance and Traffic Control Systems
car	0.020589104
cruise	0.014582608
would	0.012968431
like	0.009388691
stop	0.009103191
traffic	0.008565132
vehicle	0.007873341
human	0.00730234
driver	0.006950954
road	0.00689605
lane	0.006841146
right	0.006237203
see	0.006160337
cars	0.005907779
could	0.005841894
light	0.005765028
get	0.005688163
intersection	0.005556393
video	0.005303835
even	0.005292854
way	0.005018334
red	0.004886565
time	0.004875584
hit	0.004853622
turn	0.004842641

Word	Public needs
people	0.01655501
like	0.01078793
would	0.00902389
think	0.00770085
get	0.00637782
really	0.00595942
good	0.00575588
one	0.00574457
know	0.00533748
want	0.00497563
things	0.00457985
public	0.00448938
see	0.00446677
way	0.004365
much	0.0042293
bad	0.00413884
point	0.00400314
make	0.00391268
tesla	0.00382221
say	0.00367521
thing	0.00364128
even	0.00357344
cruise	0.00346036
actually	0.00344905
anything	0.00342643

Word	Foreign players
mobileye	0.01243205
driving	0.01194067
autonomous	0.01071223
china	0.01019629
vehicles	0.00746915
zeekr	0.00658468
technology	0.00651097
new	0.00626528
vehicle	0.00619158
also	0.00560193
self	0.00511055
us	0.00503684
chinese	0.00498771
company	0.00479116
geely	0.00454547
said	0.00442263
year	0.00437349
first	0.00407866
high	0.00405409
development	0.00353815
ford	0.00351358
system	0.00351358
2023	0.00348901
model	0.00343987
group	0.00336617

Word	Safety
cruise	0.02732235
safety	0.01359444
miles	0.01103598
vehicles	0.01004314
human	0.01002404
data	0.00897393
vehicle	0.00882118
av	0.00863025
dmv	0.00798109
drivers	0.00777106
driver	0.00593813
waymo	0.00563264
autonomous	0.0054608
report	0.00538443
incident	0.00530806
said	0.00526987
accidents	0.00517441
nhtsa	0.00517441
crash	0.00507894
public	0.00494529
driverless	0.00494529
san	0.00494529
one	0.00490711
also	0.00488801
francisco	0.00479255

Word	Big Tech players
waymo	0.016402258
cruise	0.01259624
years	0.011133382
going	0.008984405
companies	0.008634873
company	0.008091156
think	0.008039373
gm	0.0079617
would	0.007443874
year	0.007405037
market	0.006667135
money	0.00635644
google	0.006136364
like	0.005566756
scale	0.005566756
get	0.005489082
one	0.004971256
could	0.004919474
business	0.004828854
tech	0.004647615
even	0.004608778
make	0.004466376
see	0.0042463
cars	0.004129789
next	0.00383204

Word	Transit and commute
people	0.01673548
car	0.01422115
cars	0.01234043
per	0.0098261
would	0.00867957
cost	0.00835773
going	0.0081767
vehicles	0.00727154
get	0.00668822
like	0.00663793
uber	0.00655747
need	0.0065273
parking	0.00649713
city	0.0062457
vehicle	0.00607472
1	0.00595403
mile	0.00574283
use	0.00568249
transit	0.00545117
even	0.00544111
much	0.00530031
think	0.00513939
one	0.00490807
ride	0.00485779
could	0.00471698

Word	Sensor technology
data	0.0138617
lidar	0.00890934
like	0.00747277
need	0.00743497
would	0.00730896
ai	0.00667888
system	0.00569597
tesla	0.00553215
driving	0.00551955
think	0.00541874
better	0.00540614
end	0.00517931
work	0.00497769
vision	0.00496509
good	0.00471306
also	0.00467526
different	0.00462485
sensors	0.00457445
cameras	0.00447363
use	0.00430982
real	0.00429721
much	0.0041964
could	0.0041838
even	0.0041712
problem	0.00412079

Word	City trials
waymo	0.04904829
cruise	0.02547243
sf	0.01367731
one	0.00877824
service	0.00846217
like	0.0084478
driverless	0.00824667
phoenix	0.0077151
city	0.00764327
rides	0.00764327
area	0.0071979
see	0.00680999
cars	0.00680999
cities	0.00670943
ride	0.00626406
think	0.00624969
time	0.00593362
get	0.00584742
would	0.00560318
still	0.00545952
also	0.00474118
week	0.00446821
year	0.00422397
know	0.00412341
austin	0.00409467

Exhibit 6 – Top 25 words per topic

Document similarity

Our next area of interest was to explore the relationship between topics in the *news* and the sentiment online. To do this, we needed a method of assigning topics to news articles. The tool we reached for was document similarity. Our idea was the find the top 10 comments for each topic (those that scored the highest in that topic), and to calculate the similarity between a news article and each of those 10 comments. We would then take the average of those similarities and use this as a representation of the article's similarity to the topic as a whole. Completing this process for all 9 topics, we would then normalize the similarity scores, yielding a result similar to that of topic modelling.

With this methodology in mind, we chose to preprocess our comment data by selecting the set of 90 comments which represent the top comments in a topic (9 topics times 10 comments each). The '*Top 10 documents per topic*' tab contains this preprocessed dataset that was then used to compute document similarities to certain news articles, as seen in Section VI.

IV. Deep dive into select periods and news

Having defined 5 periods of time for investigation in Section III, we conducted two news scraping methodologies, collecting the article Title, date, and content in each instance:

1. Last 6 months' data: we scraped the website Motor-1, an online car news aggregator which allowed us to filter by autonomous vehicle news as shown in Exhibit 7. We furthermore scraped the Financial Times with WebScraper, using the search term 'autonomous vehicles'.

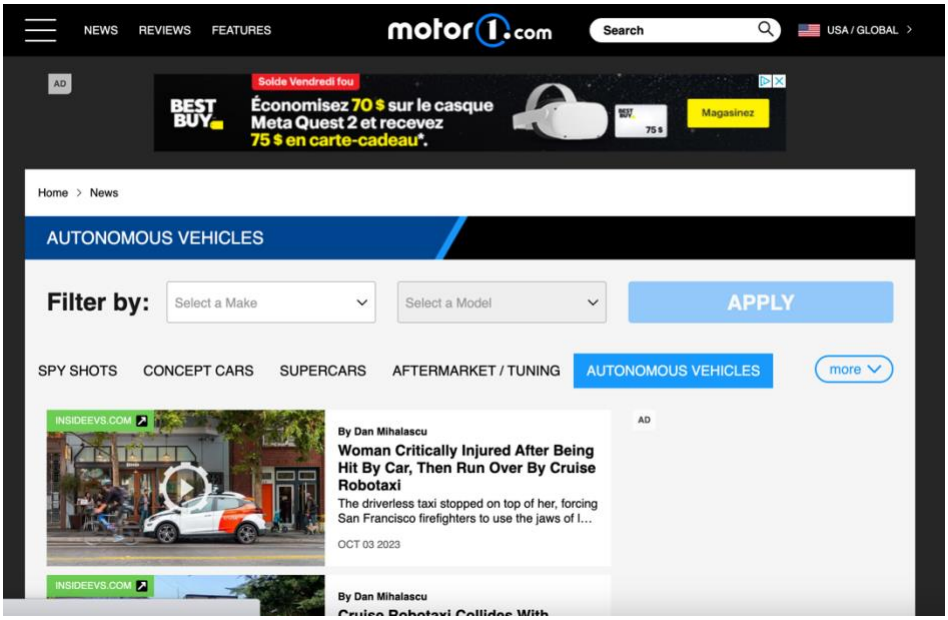

Exhibit 7 – Motor-1 AV news query

2. Four critical sentiment shift periods[5]: we scraped the Financial Times with WebScraper as it provided both a comprehensive repository of older news (easily filtered in their website advanced query) and a favorable site architecture for scraping. Exhibit 8 illustrates an example page we used for the different periods

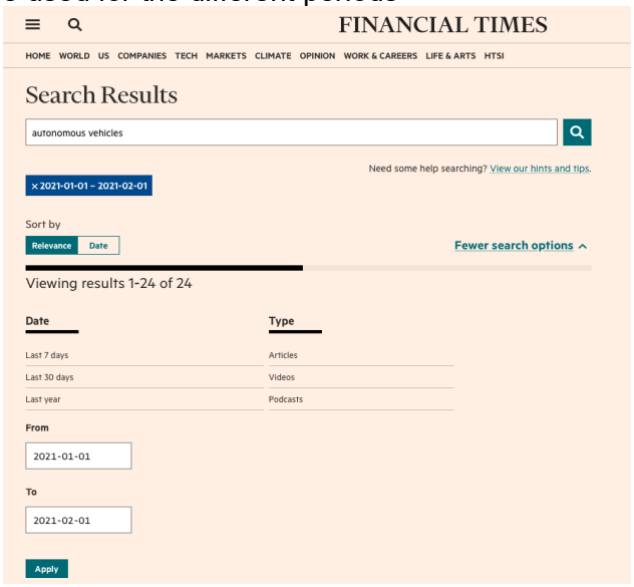

Exhibit 8 – Financial Times sample period query (here January 2021)

[5] December 2018, January 2020, March 2020, and January 2021

We scraped a total of 111 news pages (data points) with an even distribution across all the periods covered. Our news data is found in the ‘News data’ tab and the period distribution of news is displayed in Exhibit 9.

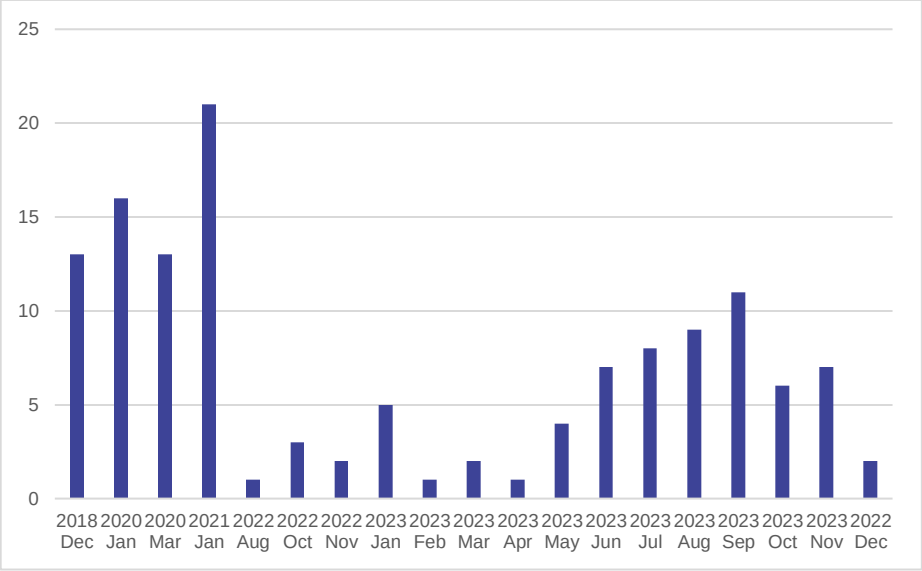

Exhibit 9 – Time distribution of scraped news data (count)

We can observe that there is a strong concentration of news on the periods defined in Section III — keeping in mind that the number of articles scraped is from only 1 or 2 sources — which is expected: periods of high sentiment shift are periods with a lot of news on the topic. As for the 2023 data, we see it is more moderate and consistent, as expected with our data being concentrated in the period.

We thus were able to compute and normalize similarity scores of the top 10 comments per topic against the news articles gathered to obtain our data points for our predictive model as seen on sheet ‘Articles & topic similarity’. We also were able to gather insights on these intermediary values by identifying minimum and maximum values as well as computing average similarity scores per topic. Some key insights we were able to conclude was that ‘Sensor Technology’, ‘Advanced Driver Assistance and Traffic Control Systems’ and ‘Foreign Players’ were the most discussed topics across articles. It is interesting to note that these topics, especially the first two, can be seen as more technical and can highlight how news articles may emphasize on discussing the low-level technicalities of AVs in more depth. We can draw a contrast with the comments left by online users, who seem to gather around topics of public interest.

V. Building a predictive model

Connecting new articles to sentiment at that time

Having scraped news articles from our periods of interest and established a methodology for assessing the topics seen in each news article (as described in Section III), we then set out to assess the relationship between the topics seen in the news and the sentiment of online discussions. To do so, we narrowed our focus to the sentiment of the comments posted on the

day of, and in the days following, the publication of the article. To do this, we looped over multi-day ranges, starting with the 2-day range including the day of publication and the day following publication, and expanded the length of our range until we obtained at least 10 comments. The rationale behind this approach was to balance a need for temporal accuracy with the need for an unbiased sample of sentiment. We found the 10-comment threshold to be effective in minimizing bias while not spanning too many days, potentially masking the impact of the news on user sentiment. The latter was especially challenging with our limited density of comments in certain date ranges (e.g. December 2018).

Building our model

Having our data preprocessed, we then set out to create a model connecting our normalized topic similarity scores for each news article, and the average user sentiment we calculated in the days following publication. As a first step, we ran a multiple-variable linear regression. The results of this regression were poor — none of the variables were significant at a 0.1 significance level, nor was the model as a whole. We then turned to machine learning techniques, specifically ridge regression. Running a regression on our dataset, we found the coefficients seen in Exhibit 10, with an RMSE of 0.04257. Given the response variable sentiment is on a scale of [-1,1], this is not very impressive and exposes some of the limitations of this dataset.

Topic	Coefficient
Big Tech players	-0.000536
Advanced Driver Assistance and Traffic Control Systems	-0.001929
Public needs	-0.000757
Foreign players	-0.000810
Safety	0.001065
FSD	0.000862
Transit and commute	-0.000264
Sensor technology	0.000682
City trials	0.001687

Exhibit 10 – Coefficients from our ridge regression model

Limitations of our model

The first major limitation of our model is the size and scope of our news article dataset. With only 111 news articles, we lack enough data points to properly generalize to out-of-sample articles. Additionally, we didn't source articles from a wide array of sources, narrowing our 'capture' of the potential influences of user sentiment. Events likely occurred, that impacted user sentiment, that were not published by FT or Motor1.

A second limitation of our model is our calculation of sentiment around the time of publication. Given four of our time periods of focus were outside of the past six months, the period containing 83% of our downloaded comments, we struggled to find representative samples of comments at specific dates of publication, forcing us to expand our search window. While expanding the timeframe used to calculate average sentiment may be beneficial in some cases, the true limitation we faced was a lack of sufficient sample sizes during some of these periods. This introduced bias and variance into our data which negatively impacted the predictive power of our model.

A third limitation of our model is the calculation of topic distributions for each news article. Given this was done using our document similarity approach, we exposed ourselves to a loss of information, as the top comments for a given topic do not define the topic, but merely represent certain aspects of said topic. A more rigorous approach would explore ways to define topics such that an out-of-sample observation, a news article, can be assigned a topic distribution from these topics in a mathematically complete way.

A fourth limitation is the choice of our model. We only explored two models, multiple variable linear regression and ridge regression. Further work can be done exploring other tools, including neural networks.

VI. Testing our model vs recent news

Cross-validation

To test our Ridge Regression, we used cross-validation to calculate an RMSE. We started by randomly splitting our data into training and test sets, with 80% of the data used to train the model and then test it on the remaining 20%. We ran this method for multiple iterations and consistently observed RMSE's of around 0.0426.

Real-world example

For a more applied exercise, we decided to test our model on a recent event of significant importance, which is the resignation of General Motors-owned Cruise's CEO on November 20th. We scraped once more for reddit comments from the same four subreddit channels and filtered for comments posted on the day of the event until now. With this new test dataset, we ran a sentiment analysis on the each of the comments and aggregated an average sentiment of -0.108: This is our true observation.

We then ran the text inside the test article through our predictive model to carve out an estimation of the mean sentiment of Reddit comments in the four subchannels following the event. The result was a predicted sentiment of -3.89×10^{4}, which quite far off the true observation and yields an error of 0.11. It is important to note that this exercise is not representative of the true performance of our model, since it does not show how the model performs over a multitude of test observations. In other words, this difference between the predicted and observed values might be a one-off outlier that would require more testing to justify as representative.

VII. Key commercial implications and strategies

As an aggregation of our multiple analyses, from our sentiment analysis to the Ridge regression, we came up with high-level strategies around Tesla's FSD superior user buy-in and perception, discrepancies in end-user communication channels, and the consumer interpretations of products that top AV makers can consider implementing to improve sentiment and increase adoption likelihood as their technology nears commercial deployment.

First, through the FSD topic we have modelled and inserted as input in our regression, we have identified a strong positive correlation with online users' sentiment. A possible implication we note is the possible superior user buy-in that Tesla is getting regarding its FSD technology. Thus, we recommend other top companies to monitor and learn from Tesla FSD's differentiating

factors as potentially attracting higher interest from their target audience. We recommend them to do so mainly by competing with Tesla on their value chain:

- Identify and replicate key FSD features and discernable design decisions (i.e., UI decisions)
- Contact the company's key suppliers to establish new contractual relationships to emulate their capabilities or learn the impending strengths and weaknesses from their supplied systems (i.e., software, sensor technology, etc.)
- Poach current and previous Tesla engineers and designers who can import industry-leading knowledge on processes and internal successes that have set the company apart.

Second, we recommend AV makers to establish a direct line of communication with their users to address the disconnect between the news shared by outlets and the way consumers actually digest and discuss received information. Through our similarity analysis between top comments per topic and relevant AVs news articles, we observed that news articles publish information with strong technical perspectives and content, whereas online users tend to focus on higher-level public perspectives that directly affect them. To address this, we advise AV makers to redesign their customer engagement experience by:

- Creating a dedicated consumer relations function with a clear mission statement to manage end-users product perception and their reception of news content
- Sharing their product roadmap and technological/testing progress regularly and more transparently with target users through social media platforms notably[6]
- Publishing regular informational and educational capsules to educate the customer on the technology, its uses cases and its impact on end-user daily routine lives. Significant learning curves still remain and companies can position themselves for success by accompanying the users along this journey, as early as possible. This continuous learning engagement aims to mitigate potential sharp losses in sentiment and buy-in resulting from confusion and mistrust following critical news (tendency to overreact).

Third, AV companies need to integrate user feedback in their product design and testing to address what they are more sensitive to – what they see and how the vehicles affect them directly. Given that we identified that online users tend to discuss more on topics that have higher relevance to them, such as '*Public Needs*' and '*City trials*', we recommend that AV companies start to take into consideration users in their product testing activities. The consumers are more prone to discuss an AV in trial without human control or an accident caused by one more than the low-level technical advancements made by a company. Their experiences interacting with the technology in their cities or by seeing it in the empirical world strongly effects online sentiment and their purchasing sensitivity. While users may be intrigued with news on capital expenditure, partnerships, or engineering decisions, their primary interest remains how this materializes near them. Potential ways to positively integrate consumer perception into testing include:

- Gifting testing sessions to specific consumers who can experience first-hand a trial while highlighting the safety of tested vehicles. This will allow regular or influential consumers to share their insights to AV makers, while improving general sentiment of users through a sense of connectivity to the progress of AVs into society.

[6] The rationale behind this choice is to use the most direct communication channels as opposed to news outlets.

- → Present solutions to the masses through public car and road shows in different cities to create both awareness and excitement, as well as educating users on safety features notably.
- → Expanding testing trials to more cities (even if they are conducted in a controlled setting). This will allow local populations to feel included in the change while circumventing local regulations that may not welcome the technology yet: the testing does not have to be on public roads (private facility). Having the technology near users is enough to embark them on the journey for adoption and incentivize them to learn more and advocate about it eventually.

VIII. Conclusion remarks

Limitations of our research

While we continue to see great potential in the connection between news and online user sentiment, we acknowledge limitations in our work. Firstly, the comment data we scraped was highly concentrated over the past 6 months. While this is not inherently bad, it limited our ability to derive insights from earlier time periods. This is the result of our budget and computational constraints, as well as the difficulty of querying for comments that are relevant to our topic. Further work can be done with a larger and more temporally distributed dataset, allowing for more statistically significant results.

We also acknowledge the limitations of our model specifically, as described in Section V.

Discussion

Our study opens the discussion for further research on the topic of autonomous vehicles and self-driving systems for automobile makers looking to justify expenditure in the field and ensure commercial success at scale. While our predictive model may not have accurately assessed sentiment upon new data insertion, we believe there is significant potential for future research building upon this paper. Furthermore, this can provide the basis for companies investing in R&D-intensive projects who are looking to evaluate users' perceptions of specific news, product features, and project developments. This can inform important decisions on product feature engineering and design as well as how to market the product to consumers.

X. References

Achal Shankar Gupta, & Sharma, S. (2022). Analysis of Public Perception of Autonomous Vehicles Based on Unlabelled Data from Twitter. 59–67. https://doi.org/10.1007/978-981-19-5331-6_7

Ding, Y., Korolov, R., Al) Wallace, W., & Wang, X. (Cara). (2021). How are sentiments on autonomous vehicles influenced? An analysis using Twitter feeds. Transportation Research Part C: Emerging Technologies, 131, 103356. https://doi.org/10.1016/j.trc.2021.103356

McKinsey and Company. (2023, January 6). Autonomous driving's future: Convenient and connected | McKinsey. www.mckinsey.com. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/autonomous-drivings-future-convenient-and-connected