What do people really think about self driving cars?

A sentiment analysis study using online social networking data to track attitudes towards autonomous vehicles and the model topics discussed about the technology.

Andrea Francu
Master of Science - Data Science Program
University of Colorado, Boulder
Boulder, CO, U.S.
andrea.francu@colorado.edu

ABSTRACT

Artificial intelligence (AI) technologies are advancing in all domains of society. One area with particularly interesting and impactful advances in AI is in transportation, with the invention of autonomous vehicles (AVs). AVs have sparked the interest of many people throughout the world and induced a variety of reactions to the growing technology, due to their potentially huge consequences. Natural language processing (NLP) tools can be used on text data from posts from Reddit, a popular online social network (OSN) platform, to analyze sentiment around AVs and model common topics discussed around this technology.

KEYWORDS

Autonomous Vehicle (AV), Online Social Network (OSN), sentiment analysis, Natural Language Processing (NLP), Machine Learning (ML), Artificial Intelligence (AI)

1 INTRODUCTION

With the rapid growth and rush for companies to achieve full autonomy of vehicles, a variety of issues and developments become talked about very quickly. Specific companies are also heavily associated with AVs and opinions about those companies change with these stories. AVs are the topic of many debates with regards to environment, safety, data, employment, and technology. This project's purpose is to look at Online Social Network data, specifically posts on the social networking site Reddit, to analysis sentiment around the AV industry as a whole, as well as with regards to specific companies that are making the most advancements.

This analysis involves processing a large amount of Reddit posts, and using topic modeling to track which topics relating to AVs are most common. It also involves performing sentiment analysis on the OSN data to track if there are positive, negative, or neutral sentiments about the AV industry in general, as well as subtopics within this industry. Subtopics may include various companies or institutions working on these technological advancements and the feelings and misconceptions about them. Other topics of interest could include false information about the capabilities of AV's, misconceptions about environmental impact, and misinformation about how far advancements have been made in this industry.

It was a bit surprising to see that the majority of subreddit pages actually received overall positive sentiment ratings, with all the fears and concerns people have about this new technology, and the impact it could have on our society, as well as safety issues. It was also interesting to see the biggest names discussed when looking at AV spaces, and learning about companies with advancements in this technology that are not as commonly discussed on other platforms or in person.

This analysis also requires looking at related work and comparing efficiency and accuracy between existing methods. Due to the limited time allowed for this study, there will be discussions of possible future work, as well as modifications and improvements that could have been made for this project.

2 RELATED WORK

There exist many different studies and tools related to various aspects of this project. Natural language processing, which includes sentiment analysis as a subdivision, is a constantly growing and adapting tool. Due to the complexity and always evolving nature of language, NLP needs to be improved all the time. Using online social networking data for data analysis also has a lot of related work and tools available. There also exist several different studies following different aspects of sentiment about AVs.

2.1 RELATED STUDIES

One related study is a scientific paper titled "How are sentiments on autonomous vehicles influenced? An analysis using Twitter feeds," written by Yue Ding et al. This paper uses machine learning to sort tweets into categories based on their tendency for positive, negative, or neutral sentiment about AVs [1]. However, this study does not look at popular topics relating to AVs or at sentiments towards specific companies.

Another paper, titled "Analyzing Self-Driving Cars on Twitter," written by Rizwan Sadiq and Mohsin Khan, looks at almost 7000 tweets and uses supervised and unsupervised learning methods to analyze sentiment about self-driving cars [2]. This analysis also does not look at common AV related topics. Both of these studies focus on sentiment analysis about the AV industry as a whole, rather than at individual aspects of it.

One analysis that did look at specific topics about AVs looked at misinformation about self-driving cars, focused on those myths that affected driver safety. People who have beliefs about AVs capabilities that are actually more advanced than what these vehicles can actually do, may rely on their cars more than is safe, putting themselves and others in danger. In this case, it is imperative to inform the public about actual capabilities, and work on eliminating these widespread myths [3]. This study's main focus is driver safety, and does not look at other misconceptions, like those relating to environmental impact, or how advanced specific companies are in this technology.

2.2 DATA MINING TOOLS

Regarding pulling OSN data into Python, there are many tools available for this. The Python Reddit API Wrapper, PRAW, allows users to access posts and various attributes of the posts. This tool is relatively easy to use and free, but limits users to the top 100 posts of any subreddit.

2.3 NLP TOOLS

When looking only at sentiment analysis methods, there are many different libraries and methods that already exist for this purpose, each with slightly varying applications and algorithms behind them. These are considered NLP techniques, and have extensive data and tuning over several years to back up their accuracy. Of course, language and communication is very complex, and always evolving, so NLP naturally has to catch up to evolve with it.

One very useful and applicable Python library is called NLTK, which stands for natural language toolkit. Within this library exist many other libraries for processing language, including tools for sentiment analysis. TextBlob is a popular and simple to use library for this. It has the capability of performing sentiment analysis, as well as extracting topics, and classification, all of which are relevant to this project. Another library within NLTK is the SentimentIntensityAnalyzer (VADER), which can return various aspects about sentiment ratings of a text. NLTK also has a toolkit called tokenize, which allows for topic modeling.

Another popular tool within the NLTK library is the VADER module, which stands for Valence Aware Dictionary and Sentiment Reasoner. This module mostly focuses on sentiment analysis tools.

Due to the vast variety of sentiment analysis tools available, the time on this project was best spent comparing between methods to see which is more applicable, rather than creating a tool similar like so many that already exist.

NLTK also has a library called tokenize, which allows for turning strings into a list of words, which can then be used for topic modeling.

3 PROPOSED WORK

The goal of this data mining project is to use OSN data to analyze sentiment about AVs, as well as using topic modeling to see the most commonly discussed subjects with regards to AVs, and the companies pioneering the technology around them. This is done by following the data mining pipeline steps, both in order and circularly as new discoveries are made. The steps of this are data understanding, data preprocessing, data warehousing, data modeling, and pattern evaluation.

The main tasks of this project include finding and organizing the OSN data, using different NLP tools to analyze sentiment around AV technology as well as finding the most commonly discussed topics about the autonomous vehicle industry, and evaluating which of these methods are more efficient and accurate for this kind of data.

3.1 DATA COLLECTION

The data used for this project consisted of Reddit posts within AV-related subreddits. Creating an account and a personal use script for free allowed for access to a client ID and client secret key, which could then be entered into the Python script using the Python Reddit API Wrapper (PRAW) to access subreddit threads. This was done by entering the given credentials into a PRAW object, and then searching for a specific subreddit. Previous research was also done to find the relevant subreddit threads for this project and the exact naming of them. The subreddit threads used for this project included: "AutonomousVehicles," "SelfDrivingCars," SelfDrivingCarsLie," "cars," "SelfdrivingcarsWCGW" (what could go wrong), and then various subreddits about the leading companies, found from topic modeling in the previous threads. These subreddit threads included: "teslamotors," "ford," "waymo," "GeneralMotors," "Zoox_inc," and "Toyota."

Once these subreddits were accessed, lists were made with attributes including the string of the post, the ID of the post, and the number of "upvotes" (likes) within each thread which was then put into a data frame for each thread.

3.2 DATA UNDERSTANDING

Understanding the data used included researching through Reddit for applicable threads, as well as looking through those threads to see common topics. This was an iterative process as originally only the first group of subreddit threads listed were included. But after performing topic modeling on those threads, it became clear that there were a few companies that were talked about a lot, and so that is how those threads were chosen to be added to the data sourcing.

There was also some data understanding required when looking at data to use for evaluating the NLP tools used. In order to test the limits of the sentiment ratings, data that was very positive, very negative, and neutral had to be used and ensured to be near those limits.

This also included understanding the AV space, including companies with the most advancements and funding in this area, as well as typical sentiment and topics discussed around the technology, and a basic understanding of how the technology even works.

3.3 DATA PREPROCESSING

Preprocessing the data included only selecting necessary attributes, including the string of the text, ID, and number of upvotes. Since PRAW gave a limit of 100 posts from each subreddit, not much data warehousing was necessary to organize the data.

3.4 DATA MODELING

Modeling this data consisted of using already available NLP tools for sentiment analysis and topic modeling. One popular NLP tool is the NLTK library, previously mentioned. This toolkit has many different libraries available for a variety to NLP tools.

The TextBlob tool within this was used to looping through each post pulled from the subreddit, and assign a sentiment score. A sentiment score is a number between -1 (negative sentiment) and 1 (positive sentiment).TextBlob also has a method that returns the subjectivity of that rating, which can be very useful for evaluating the method. The other NLTK tool used, VADER, also gives a sentiment score between -1 and 1. These were the two main methods used to analyze sentiment and evaluate between the two.

Tokenize was then used to turn strings into a list of words. Then the frequency of each word was counted and the list was sorted to show the most commonly used words, for topic modeling.

4 EVALUATION

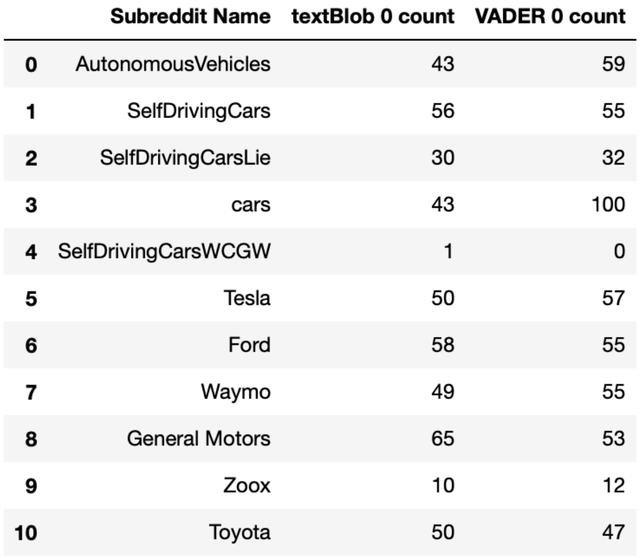
Evaluation of this analysis requires comparing different possible methods for sentiment analysis. The main evaluation metrics are effectiveness and efficiency.

With the multitude of methods available for conducting sentiment analysis, the hardest task is often finding the best one for a specific application. Finding the best method can be difficult, since language and communications are so complex, and sentiment is a subjective topic. For this project, the Reddit posts, and their sentiment ratings from different tools, were stored in a data frame. Since sentiment ratings are numbers from -1 (negative sentiment) to 1 (positive sentiment), various numerical comparisons and statistics can be drawn from these scores.

One way to evaluate which method is more accurate is to look at the posts that received neutral ratings. This comparison requires manually looking through some of these posts, to see whether or not there really is sentiment there when the tool is reporting neutrality, or if one tool is more likely to give a rating even when there isn't any sentiment. If it is clear that these neutral ratings are skewing one way, then one metric for comparing methods can simply be which one has more (or less) neutral ratings.

For all the subreddit channels analyzed, the TextBlob method gave 454 total zero scores, whereas the VADER method gave 526 total zero scores. Looking through some of the posts, it did seem like VADER was typically more willing to give a sentiment score even if there was no sentiment. However, it also seemed like TextBlob was less willing to give a score even if there was sentiment. So depending on the type of analysis being done, each method could be better for different situations. If it were a situation where getting some incorrect data could costs people's lives, then it would probably be better to go with the safer option of TextBlob analyzer. However, for purposes like this project where there isn't that much data to work with, it may be better to go with a tool that's more willing to give a score, so there are more actual scores and data to work with, like with VADER.

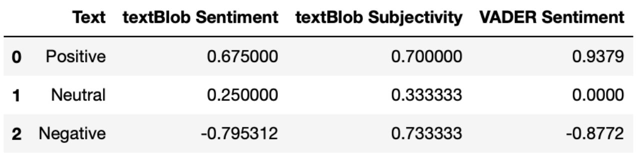
Below is the number of zero scores given for each subreddit thread by the two different tools.



	Subreddit Name	textBlob 0 count	VADER 0 count
0	AutonomousVehicles	43	59
1	SelfDrivingCars	56	55
2	SelfDrivingCarsLie	30	32
3	cars	43	100
4	SelfDrivingCarsWCGW	1	0
5	Tesla	50	57
6	Ford	58	55
7	Waymo	49	55
8	General Motors	65	53
9	Zoox	10	12
10	Toyota	50	47

Another way to compare between scores is by looking at the average sentiment rating for a whole subreddit. Manually sorting through posts to see what kind of rating seems more accurate is one way to use this metric. Another way is to use the different methods on texts that are very positively or negatively skewed, and see which method gets closest to the positive or negative expected rating, as well as comparing to a very neutral text.

This analysis was carried out and the results are shown in the table below.



	Text	textBlob Sentiment	textBlob Subjectivity	VADER Sentiment
0	Positive	0.675000	0.700000	0.9379
1	Neutral	0.250000	0.333333	0.0000
2	Negative	-0.795312	0.733333	-0.8772

VADER seemed to be the more accurate method for all three cases. For the very positive text, it gave a sentiment score of 0.94, whereas TextBlob gave it a score of 0.68. For the neutral text, TextBlob gave it a positive score, as well as a subjectivity score of 0.33, where VADER gave it a score of 0. And for the negative text, VADER gave a score of -0.88 while TextBlob gave a score of -0.79. This analysis proved that VADER was more accurate in rating positive, negative, and neutral scores.

5 DISCUSSION

5.1 TIMELINE

With no prior experience in sentiment analysis, topic modeling, or OSN mining, making a feasible plan for this project proved a bit difficult. After completing some of the work, certain parts proved more or less challenging than expected.

This project was completed over the span of five weeks. The first week was for general research and the proposal. This was the time to find topics of interest, narrow down research, and look at related work.

The second week was originally planned to focus on the data and visualization. This would have included sourcing the data, understanding the data, and preprocessing the data to make it easier to manage. The visualization aspect would include general statistical analysis, plotting variables of interest, and making a dashboard of data visualizations.

Week three would have focused on topic modeling and sentiment analysis. This would have begun with manually analyzing sentiment on chose topics, but later be automated for larger data sets if time allows. However, weeks two and three were switched around a bit. Week two was still spent sourcing the data, understanding it, and preprocessing it, but instead of going into the visualization aspect from there, the project began focusing on sentiment analysis. This is because sentiment analysis is so core to the project, and with no prior personal experience in this area, it was strategic to begin with this task to ensure feasibility. Going straight into sentiment analysis proved to not be too difficult, since there are such a wide range of tools already available. This showed that more of the work would have to be done in researching and evaluating methods, rather than actually taking the time to code from scratch a tool that would likely not be as accurate as already existing ones, in the allotted time.

Week three was then slated to focus on visualization and statistical analysis of the methods for evaluation. In addition to this, more subreddits were looked at for sentiment analysis, given the 100 post maximum data scraping allowed per subreddit. Given that the code for mining a subreddit thread and adding it to the data frame was already written, adding more of them was relatively simple. But as results came in from some of the tools used, this paved the way for areas that should be more focused on.

Week four was slated to be for model evaluation. This was the time to compare between different approaches as well as with related work. However, since this was moved to be done a week earlier, week four was then planned to be for looking at topic modeling.

Week five was for working on the final report and slides, and working on the final presentation, practicing it, and recording the presentation.

The timeline for this project did change throughout and some of the steps were iterative and done over multiple weeks, but having this general outline definitely made the project more manageable and organized.

5.2 DATA MINING

The original plan for this project was to use data from Twitter, specifically, tweets including specific words (like “autonomous vehicle,” “self-driving car,” etc). The first challenge that arose was in mining this OSN data. A large part of sentiment analysis research on OSN data is done on tweets from Twitter. Because of this, there are many available tools to aid in mining tweets. However, Twitter has varying levels of authentication based on the type of account a user has, and based on how much they are willing to pay for. The first attempt of this project was to see what data could be accessed with the free, base level account. From researching online, it seemed like it would be possible to access tweets with this account and import them into Python. However, after attempts with several different methods, errors of the same sort kept popping up, relating to the Twitter API access levels that did and didn’t allow for this mining. After some more research, it seemed like Twitter had recently changed their policies on what could be access with which type of account, and no tweets could be mined with the base account.

This then led to looking at what data could be accessed for free from Reddit, a similar OSN platform with a different layout. This proved to be much simpler. With the creation of a personal use script application account, users can access the 100 top posts in any subreddit. Though this isn’t too much data, multiple subreddits can be accessed, allowing for a larger compilation. This was done with PRAW, the Python Reddit API Wrapper mentioned previously.

Using this Python package, the top 100 posts in a subreddit, along with the ID of the post and the number of upvotes, can be pulled into Python, and then is organized into a data frame.

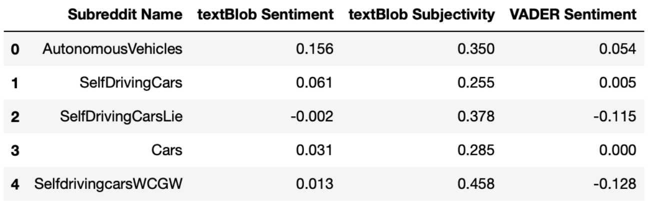
The original plan for this project, when looking at tweets rather than Reddit posts, was to use keywords as queries. The keywords that would have been used to find these posts would have included “autonomous,” “self-driving,” and “driverless.” Since Reddit has a different format, where there are subreddit topic pages and people post within that page for the topic, a lot of the topic searching part of the work is already done. Further topic modeling within these subreddits could be done to track misinformation, or comparing between companies. As a starting point for OSN data, the top 100 posts in the subreddits titled “AutonomousVehicles,” “SelfDrivingCars,” SelfDrivingCarsLie,” “cars,” “SelfdrivingcarsWCGW” (what could go wrong) were looked at. Each subreddit was mined for top 100 posts, with the associated ID’s, and the number of upvotes.

After some of the results came in from topic modeling, it became clear that a few specific companies were being talked about the most in many of the threads. Tesla was the first or second most common word used in four of the five subreddits. These words most discussed became the basis for the further subreddits searched. These included various AV/car manufacturers including “teslamotors,” “ford,” “waymo,” “GeneralMotors,” “Zoox_inc,” and “Toyota.”

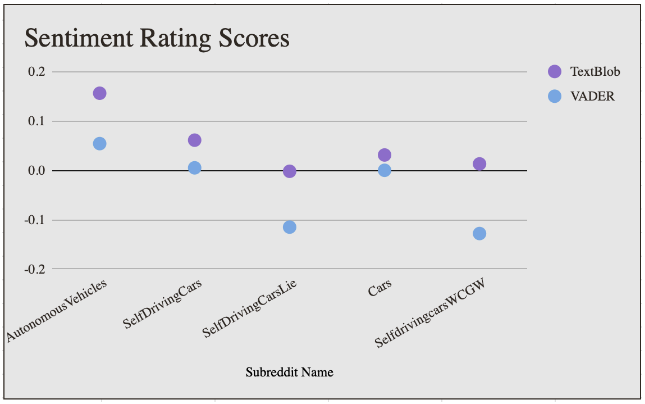
5.3 SENTIMENT ANALYSIS

The original plan for this project was to create a sentiment analysis tool, by looking for typically positive and negative words and giving a score to each post. After researching sentiment analysis more, it became clear there exist many tools for this already, that have used machine learning, perfected algorithms, and huge amounts of data. Rather than spending time trying to create a tool that already exists, the new plan was to use already existing methods and compare between them to find the most optimal.

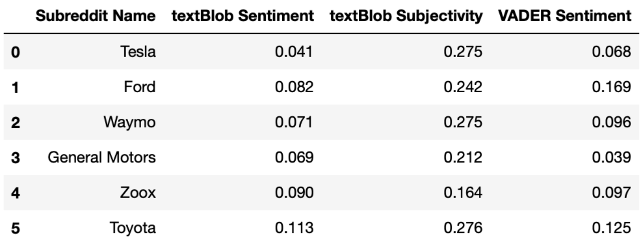
So using the two packages discussed earlier for sentiment analysis, TextBlob, and VADER, both within the NLTK library, each post within each subreddit was given a sentiment score, as well as a subjectivity score by TextBlob. The subjectivity score gives a rating of how subjective the text inputted is. After all the posts were given these scores, an average was taken for each subreddit. These averages were put into data frames, seen below. The first table and show the ratings related to AV's.



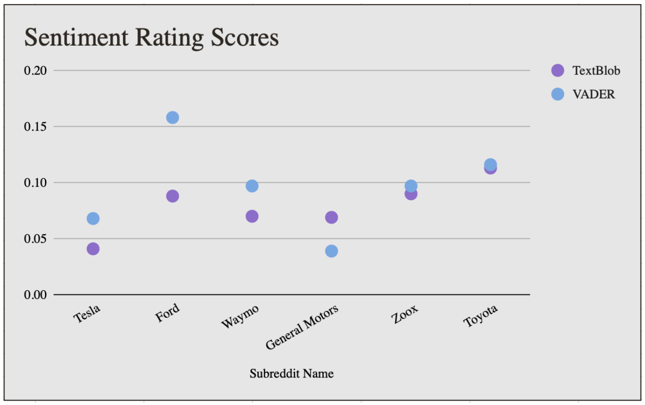
	Subreddit Name	textBlob Sentiment	textBlob Subjectivity	VADER Sentiment
0	AutonomousVehicles	0.156	0.350	0.054
1	SelfDrivingCars	0.061	0.255	0.005
2	SelfDrivingCarsLie	-0.002	0.378	-0.115
3	Cars	0.031	0.285	0.000
4	SelfdrivingcarsWCGW	0.013	0.458	-0.128



This second data frame includes the ratings of the various AV/car company's subreddits.



	Subreddit Name	textBlob Sentiment	textBlob Subjectivity	VADER Sentiment
0	Tesla	0.041	0.275	0.068
1	Ford	0.082	0.242	0.169
2	Waymo	0.071	0.275	0.096
3	General Motors	0.069	0.212	0.039
4	Zoox	0.090	0.164	0.097
5	Toyota	0.113	0.276	0.125



It was surprising to see that only one of the eleven subreddit threads was given a negative score by both tools, and one other one got a negative score from the VADER tool. The page that got a negative score from both tools is an inherently negative subreddit, titled "SelfDrivingCarsLie," so it is not surprising that it would have negative sentiment. TextBlob did give it a rating very close to zero (-0.002), while VADER gave it a score closer to -0.1. The page that got a negative score only from VADER also got a TextBlob score very close to zero. This page had much less traffic than the others, with only three total posts.

Overall, none of the subreddit communities got sentiment ratings higher than 0.158, showing that none of the pages were overwhelmingly positive.

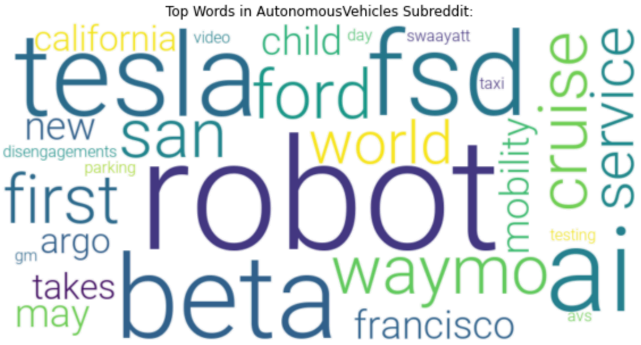
VADER pretty consistently gave higher scores than TextBlob, in all cases except General Motors, a subreddit that is arguably pretty neutral.

When evaluating the sentiment analysis tools, research had to be done to find very positive, very negative, and very neutral subreddit pages. After a bit of research, the page that was settled on for negativity was titled "AITAH" (am I the asshole), the page for neutrality was "Mildly Interesting" and the page for positivity was titled "positivity." However, none of these pages had scores very close to -1 or 1. So, going to a simpler method, instead of using OSN data to test the sentiment analysis tools, text was inputted directly into Python. Using very positive text, very negative text, and very neutral text, the methods were then tests on each text and this gave the much clearer results discussed earlier.

5.4 TOPIC MODELING

Topic modeling proved to be a fairly straightforward task, since there are many tools readily available. The text of all the posts in a subreddit page were merged into one body. Then, using the tokenize library in NLTK, each word was tokenized into it's own word. After this, each word was counted for the number of times it appeared, and then the list of words with their word counts sorted in descending order. The first time this was done, the most common words that showed up were words like "as," "the," "so," "a," along with common punctuation. Looking at solutions for this online, NLTK has a list of "stop words" that include all those filler words that people typically don't want included when topic modeling. The texts were then reanalyzed without including any stop words or punctuation. This led to a lot of the top words being words like “autonomous,” “car,” “vehicle,” etc. which also did not really contribute any data since these would naturally be discussed in these pages. So after removing these words too, a final list of the top 30 words in each page was created. These were then made into visualizations using word clouds, from the wordcloud library. These visualizations helped drive a lot of the remaining research in the project, including the specific companies to look at discussions about.

A main subreddit page looked at was the “AutonomousVehicles” one. This included a lot of the big companies working in and funding AV development. These companies include Tesla, Waymo, General Motors, and Ford.



Another topic that came up in many of the pages was San Francisco. This makes sense when looking at all the advancements in self driving car technology and legislation in the city. There are driverless taxi’s all over San Francisco now, and many people comment on incidents that happen with these vehicles.

5.5 FUTURE WORK

Given the short amount of time allotted for this project, there are many improvements and further work that could be made.

One area that could be a benefit for research to look into is the discussion of misinformation or myths surrounding AVs. Many of these myths affect driver safety, the environment, investment funds, and many more. Sentiment is also heavily influenced by incorrect information that circulates the internet. These sentiments can then affect the future of the industry, which can impact so much of society.

Another aspect of the project that could be expanded on is the amount of data looked at. While many different pages were analyzed, PRAW did only allow for the top 100 posts in any page, limiting the amount of data that came in. This also only accounted for actual posts, but comments within that post could not be imported. Furthermore, the data from these posts only included text data, even though Reddit has plenty of image and video data, which could definitely be pertinent to sentiment around and topics discussed about AVs. This data could be very useful if analyzed correctly.

Furthermore, it could be interesting to expand this work to other OSN sites, like Twitter, Instagram, TikTok, etc. and compare between sentiment on different sites. Some OSN sites are known for being overall more positive or more negative than others, so seeing the extremities of sentiment around AVs would be helped by expanding to more platforms.

6 CONCLUSION

With the quickly growing advancements in technology and AI come huge developments in the race for autonomous vehicles. With any new technology, especially ones that can potentially affect millions of people, and have many safety concerns, come worries and complaints. However, from online forums on the social networking site Reddit, many people are discussing these advancements in positive ways. The companies and places with most advancements and funding relating to AV projects are the ones discussed the most, across many different subreddit pages. Sentiment analysis is a constantly growing domain, changing as language and communications change, and using this along with topic modeling can be very useful for projections about the future of the AV industry.

REFERENCES

[1] Yue Ding, Rostyslav Korolov, William (Al) Wallace, Xiaokun (Cara) Wang, How are sentiments on autonomous vehicles influenced? An analysis using Twitter feeds, Transportation Research Part C: Emerging Technologies, Volume 131, 2021, 103356, ISSN 0968-090X,https://doi.org/10.1016/j.trc.2021.103356.

[2] Sadiq, Rizwan, and Mohsin Khan, “Analyzing Self-Driving Cars on Twitter.” arXiv.Org, 5 Apr. 2018, https://arxiv.org/pdf/1804.04058.pdf

[3] McDonald, Tony, et al. “Data Mining Twitter to Improve Automated Vehicle Safety.” Safe-D Safety Through Disruption, 1 Feb. 2021, https://www.google.com/url?q=https://rosap.ntl.bts.gov/view/dot/56364/dot_56364_DS1.pdf&sa=D&source=editors&ust=1694397137262286&usg=AOvVaw13-Te4yU938WsH76hIsgfd