Cause-and-Effect Analysis on Autonomous Vehicle Disengagement with NLP Deep Transfer Learning: A Scalable End-to-End Pipeline Approach Using the California DMV Dataset

Yangtao Zhang

April 30, 2021

Abstract

The advancement in Machine Learning and Artificial Intelligence is promoting the testing and deployment of AVs on public roads. The California Department of Motor Vehicles (CA DMV) has launched the Autonomous Vehicle Tester Program, which collects and releases reports related to Autonomous Vehicle Disengagement (AVD) from autonomous driving. Understanding the causes of AVD is critical to improve the safety and stability of the AV system, and provide guidance for AV testing and deployment. In this work, a scalable end-to-end pipeline is constructed to collect, process, model, and analyze the disengagement reports released from 2014 to 2020 using natural language processing deep transfer learning. The analysis of disengagement data using taxonomy, visualization and statistical tests reveals trends of AV testing, categorized cause frequency, and significant relationships between causes and effects of AVD. We found that (1) manufacturers tested AVs intensively during the Spring and/or Winter. (2) test drivers initiated more than 80% of the disengagement while more than 75% of the disengagement were led by the errors in perception, localization & mapping, planning and control of the AV system itself (3)there was a significant relationship between the initiator of AVD and the cause category. This study serves as a successful practice of deep transfer learning using pre-trained models and generates a consolidated disengagement database allowing further investigation for other researchers.

1 Introduction

The advancement of Machine Learning and Artificial Intelligence and artificial intelligence is bringing Autonomous Vehicles (AVs) closer to the roads with potential benefits to save people's lives, improve efficiency, reduce energy consumption, and so on. According to the definition from the Society of Automotive Engineers (SAE) [1], there are six levels of diving automation from Level 0 (No Driving Automation) to Level 5 (Full Driving Automation). While manufacturers are making the effort to deliver Level 5 AVs in the future, most AVs driving or testing on the public roads nowadays are SAE Level 2 (partial automation), Level 3 (conditional automation), and Level 4 (high automation

in geofenced areas) vehicles. Level 2 automation requires human drivers to stay alert and intervene whenever necessary and Level 3 automation requires drivers to take over control whenever requested [2–6].

The takeover control transition from the AV to the human driver is known as the Autonomous Vehicle Disengagement (AVD). Formally, the California Department of Motor Vehicles (CA DMV) defines AVD as "a deactivation of the autonomous mode when a failure of the autonomous technology is detected or when the safe operation of the vehicle requires that the autonomous vehicle test driver disengage the autonomous mode and take immediate manual control of the vehicle [7]." Understanding the causes and effects of AVD is essential to evaluate the safety and stability of current AVs testing on the roads and to provide guidelines on the regulations for AV testing and deployment.

In September 2014, the CV DMV initiated the Autonomous Vehicle Tester Program which allows permit-holding manufacturers to test AVs with a human driver in the driver seat on public infrastructure [8]. The program also requires permit holders to track and submit disengagement reports when "their vehicles need to disengage from the autonomous mode during tests" and collision reports for "every collision involving one of their vehicles" [8]. Since the establishment of the program, the CA DMV has been releasing disengagement reports and collision reports to the public on a yearly basis. These reports consisting of thorough and specific raw data are ideal data sources for investigating AVD.

By making use of the CA DMV disengagement reports, this thesis research makes the following contributions: First, in order to analyze the cause-and-effect relationships of AVD utilizing the disengagement reports and collision reports from 2014 to 2020, we built a scalable end-to-end data pipeline, which can process and analyze historical data as well as new incoming reports with satisfactory performance. During the development of the data pipeline, a comprehensive and consolidated database was generated. Second, in order to achieve the optimal performance with the limited amount of data collected, we proposed an ELECTRA-based natural language processing (NLP) deep learning model that was first pre-trained on a large scale natural language corpus (e.g., Wikepedia, Google News), then fine-tuned on a task-specific dataset (SemEval-2010 Task 8 [9]) extended by Li et al. [10] to include embedded causality [11], and finally post-trained on the CA DMV dataset. Third, we developed a taxonomy of the causes and effects of AVD with a focus on how AV system worked and interacted with other factors.

2 Related Work

According to the disengagement reports released by CA DMV, the initiator of disengagement can either be the AV System when it fails to execute due to technical issues and thus requests the human driver to take over control, or the test drivers when they feel uncomfortable with or do not trust the AV system and thus take over control proactively. Researchers have conducted studies on both system-initiated and driver-initiated takeovers.

However, previous studies focused on the exploration of the disengagement using traditional statistical and machine learning models. Favarò et al. [12] analyzed the contributory factors, disengagement frequencies of AVD with taxonomy and statistical

visualization and provided a comprehensive overview of the disengagement dataset and trends of reporting, which highlighted the limitations of the current regulation. Boggs et al. [13] used logistic regression and taxonomy to identify and quantify who, what, when, where and why (5 Ws) of AVD, and they found the disengagement initiator was linked to contributing factors derived from 5 Ws. Wang et al. [14] used multiple statistical modeling approaches and classification tree to quantitatively investigate the underlying causes of AVD, and found that lacking a certain number of sensors significantly induced AVD. Alambeigi et al. [15] used probabilistic topic modeling to examine the open-ended crash narratives and identified themes for further analysis. Their findings emphasized the safety concerns with transitions of AV system to human control.

In the past years, studies also have been conducted to investigate driver behaviors related to AVD, and their findings could help better understand the human factors contributing to AVD. Vindhya et al. [16] used cluster analysis and multinomial logistic regression to build a statistical model, in which perceptual cues were utilized to capture collision-avoidance behaviors. The results showed the mode and timing of an alert from the vehicle system did not influence the driver's behavior. Their driver model could help guide the design of more sophisticated vehicle automation systems. Markkula et al. [17] created simulation-ready human behavior models to reproduce qualitative patterns of important scenarios like "an AV handing over control to a human driver in a critical rear-end situation". With computer simulations, their models allowed optimization of AV impacts on safety.

In addition to the above-mentioned studies, Banerjee et al. [18] applied the data pipeline method to process and analyze data from the system's perspective with over a million miles of domain data. They found that the AVs' machine-learning-based system for perception, decision, and control were the primary causes of AVD, and human operators of AVs had to stay as alert as drivers of non-AVs. However, they did not utilize the complete disengagement reports released by CA DMV and further reserach is needed. What's more, their data pipeline was not scalable and could not be applied to analyze new incoming disengagement reports to benefit future analysis.

Considering all the advantages and limitations of previous studies, this study created a scalable end-to-end pipeline using NLP deep transfer learning and built a taxonomy of causes of AVD with the most thorough data to date from CA DMV.

3 Methodology

A scalable end-to-end pipeline (Fig. 1) was constructed to collect, process, model, and analyze the disengagement reports with high efficiency and accuracy. The pipeline consisted of four stages. At stage one, it collected multi-format disengagement reports from the CA DMV disengagement database and classified them into corresponding years. At stage two, it used Optical Character Recognition (OCR) to extract information from the PDF format reports and exported to CSV files. Then those CSV files were filtered, cleaned, parsed, and finally labeled by human workers. At stage three, it applied NLP deep transfer learning to train and evaluate pre-trained models and the best model was used for prediction on incoming disengagement reports. At stage four, it analyzed the disengagement database by creating a taxonomy, summarizing results in visualization,

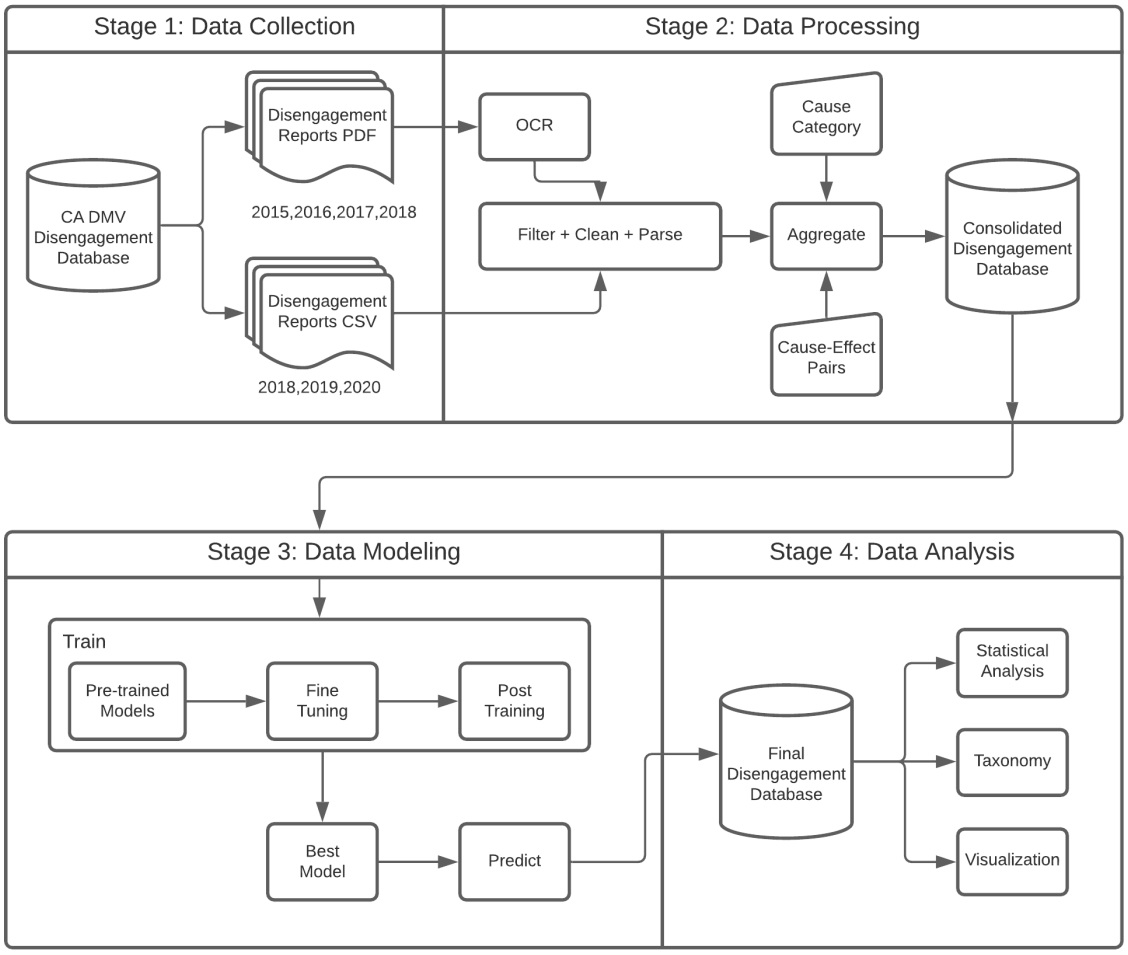

Figure 1: Overview of the data pipeline from stage 1 to stage 4

and statistics. The pipeline also generated a consolidated database which can be utilized by other researchers interested in this topic.

3.1 Data Collection

The raw disengagement data were retrieved from the CA DMV's disengagement database, which included the disengagement reports collected from manufacturers participating in the Autonomous Vehicle Tester Program on a yearly basis. For this study, the disengagement reports released from 2014 to 2020 were used as the initial data for processing and modeling in this thesis. However, the proposed method can be updated easily with new data.

The format requirement of the disengagement reports has changed over time as shown in Table 1, which also reflected the hard work done by the CA DMV to regulate the tester program. From 2014 to 2017, the CA DMV put forward general requirements for the disengagement reports, based on which different manufacturers developed their own formats for disengagement reports as long as they included the number of miles for the autonomous distance and the number of incidents for the AVD. Some manufactures, like

GM Cruise, only recorded date, time, and causal factors, while others, such as Nissan, recorded date, time, testing conditions, locations, weather/road conditions, elapsed time and type. Inconsistent formats led to inconsistent data fields, which made it difficult to compare the performances of AVs across manufacturers and to identify factors important to AVD. The CA DMV realised the lack of consistent formats, and in 2018 it introduced a revised template used for the annual report of autonomous vehicle disengagement. The standardized annual report template ensured the integrity of the disengagement data. However, it was still difficult to analyze disengagement data at scale since extracting information from reports in the PDF format was time-consuming. Thus, later in 2019, the CA DMV designed a spreadsheet to help manufacturers to put together all the disengagement logs. Once the CA DMV had collected the spreadsheets from the manufacturers, it combined all the records together and released two final CSV files - Autonomous Vehicle Disengagement Reports and Autonomous Mileage Reports, which were easy to process and distribute to the public. For each record, the final Autonomous Vehicle Disengagement Reports contained 9 fields.

With the disengagement reports in the last 7 years (2014 - 2020) collected and inspected, manufacturer, date, initiator, location and description were selected as the fields for the consolidated disengagement database and were used for further analysis. Table 2 shows the sample data and format in the database. N/A was introduced as the placeholder for missing values to ensure the integrity of the database.

Table 1: How format of disengagement reports changed over time

Years	Formats
2014, 2015, 2016, 2017	Manufacturers had their own formats
2018	The CA DMV introduced standard PDF template for annual reports
2018, 2019, 2020	The CA DMV released consolidated CSV files for disengagement reports

Table 2: Sample data and format for collected reports

Manufacturer	Date	Initiator	Location	Description
EasyMile	11/30/2020	AV System	Street	A collision hazard in the environment ahead was detected by the software, which triggered an Estop
Apple	06/19/2019	AV System	Street	Motion planning timed out
Uber	03/01/2018	Test Driver	Street	Precautionary Takeover or Operator Discretion
Waymo	09/01/2017	N/A	Highway	Disengage for a software discrepancy
Tesla	10/15/2016	AV System	Freeway	Follower Output Invalid
Volkswagen	06/12/2015	N/A	N/A	Planner not ready

3.2 Data Processing

The collected data with various formats required further data processing. On the one hand, OCR was used to recognize the text from the PDF files. On the other hand,

human insight was required to better understand the casual-effect relationships and cause categories, which was aggregated to generate the ground-truth training data for modeling cause-and-effect relationships.

3.2.1 OCR

An OCR pipeline (Fig. 2) based on OpenCV [19], Tesseract [20] and PyImageSearch [21] was built to extract texts from the PDF files, complie them, and export them to CSV files. As the PDF files submitted by various manufacturers were scanned, they were subject to random scaling, rotation, and skew. To fix this problem, a standard template was used as the reference to adjust other scanned disengagement reports to generate the de-skewed versions of images. The bounding boxes were then identified and the text within them were extracted. Next, raw texts were cleaned, filtered, and finally exported as CSV files.

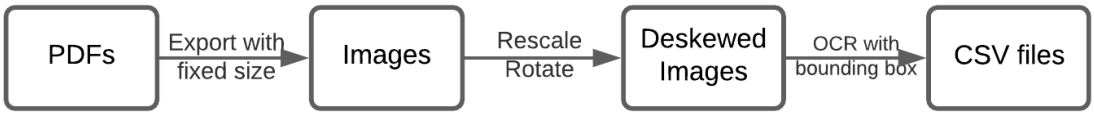

Figure 2: Overview of the OCR pipeline

With the OCR pipeline, PDF format disengagement reports from 2014 to 2018 were successfully converted into CSV files, which were combined with CSV files released by CA DMV from 2018 (see Table 1 as part of the reports in 2018 were in CSV format) to 2020 to reach a consistent format.

3.2.2 Human Insight

Two types of human insights were required for this study: cause-and-effect relationship labeling and cause categorization. The former was used to identify the causes from the disengagement reports for training the NLP deep learning models and the latter contributed to classifying the causes into proper categories.

For cause-and-effect relationships, both simple causality and embedded causality were considered in this thesis. Following the standard cause-and-effect tagging format, the Inside, Outside, Beginning (IOB) notation was used to label the tokens. To be more specific, suffix C - Cause, E - Effect and CE - Embedded Cause were added to the IOB notations. For example, B-C represents the beginning of a cause token while I-E represents the interior of an effect token. Based on human insights, each word in the description of disengagement was tagged with labels and categories and an example is shown in Table 3.

For the cause category, 3 main categories (AV System, Human Factors, Environmental Factors Others) and 9 subcategories (Perception, Localization & Mapping, Planning, Control, System General, AV Driver, Other Driver & Vehicle, Environment, Other) were derived from the conceptual organization highlighting trust development produced by Schaefer et al. [22], the AV hierarchical control structure drawn by Banerjee et al. [18],

and the taxonomy developed by Boggs et al. [13]. Each cause was placed into the most appropriate category by three workers through ground truth aggregation.

Table 3: Sample sentence for IOB label and cause category. (B-E represents the beginning of an effect, I-E represents the interior of an effect. B-C represents the beginning of a cause, I-C represents the interior of a cause. 2 means the cause belongs to the planning category)

Words	**Label**	**Category**
driver	B-E	
disengagement	I-E	
due	O	
to	O	
planning	B-C	2 - planning
discrepancy	I-C	2 - planning
in	O	
the	O	
determination	O	
of	O	
autonomous	O	
vehicle	O	
speed	O	

3.2.3 Ground Truth by Human Labeling

In this study, three student workers, who had at least 3 months experience in this field, provided the human insights on labeling the cause-and-effect relationships and cause categorization. While different workers provided their insights, it was necessary to aggregate those insights and to determine the ground truth. The framework CrowdTruth [23] was used for ground truth aggregation. It not only selected the most reliable labels among workers but also offered useful metrics, including Worker Quality Score (WQS) and Annotation Quality Score (AQS) to evaluate the performance of workers as well as the quality of their annotations.

As shown in Table 5 and Table 4, the WQS and AQS of the ground truth aggregation were higher than 0.9, which provided convincing support for the quality of the human annotation.

Table 4: Annotation quality score for 8 labels

Label	**AQS**
O	0.9942
B-C	0.9257
I-C	0.9352
B-E	0.9461
I-E	0.9184
B-CE	0.9042
I-CE	0.9331

Table 5: Worker quality score for 3 workers

Worker Id	WQS
0	0.9802
1	0.9751
2	0.9845

3.2.4 Consolidated Database Summary

Apart from OCR and human labeling, the disengagement reports with cause description less than 5 words were also filtered to reduce the noise existed in the raw data. What's more, as discovered by Boggs, et al [24], "Apple and Uber lack a variation among human-initiated disengagements" in terms of cause description and those disengagement reports replicated the same information hundreds of thousands times. Therefore, those records were excluded as well. With the completion of data processing, a consolidated database (Fig 3) was generated. It contained information of four entities - report, description, cause, and word. The four entities were weaved together logically - reports had descriptions, causes existed in descriptions, and causes were made of words. Table 6 lists the number of filtered disengagement reports of each year submitted by manufacturers in the database and Table 7 describes the number of unique entities in the database. It was efficient to filter certain entities and conduct further data analysis by utilizing the query of the consolidated database.

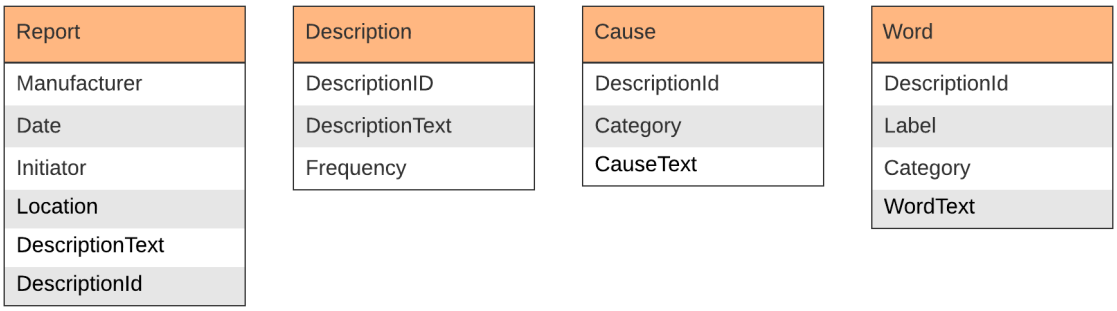

Figure 3: Overview of entities in the database

Table 6: The number filtered disengagement reports of each year in the consolidated database

Year	Count
2014	40
2015	137
2016	317
2017	878
2018	2394
2019	7363
2020	3153

Table 7: The number of unique entities in the consolidated database

Entity	Count
Report	14282
Description	1036
Cause	377

3.3 NLP-based Deep Learning Model

We proposed an NLP deep learning model based on ELECTRA [25] using transfer learning. First, ELECTRA pre-trained by google on Wikipedia and BooksCorpus consisting of 3.3 Billion tokens was imported. Second, ELECTRA was fine-tuned on the SemEval-2010 Task 8 [9]) to gain task-specific knowledge. Third, ELECTRA was post-trained on the consolidated disengagement dataset created in stage two of the data pipeline to further improve the performance of the pre-trained models. In order to compare with other popular NLP deep learning models, BERT, DistilBERT, and XLENT were also included. These models were evaluated based on weighted F-1 score which calculated the average among labels weighted by support to handle label imbalance existed in the training data, and the cost of computational resources to select the best model. The best model can be used to extract cause-and-effect relationships and classify causes into different categories for the future incoming large amount of disengagement reports, which will save the future manual work.

3.3.1 Transfer Learning

As Torrey and Shavlik [26] noted, "Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned". For example, the knowledge gained while learning to classify Wikipedia texts can help tackle legal text classification problems. With the help of transfer learning, the data modeling did not start from scratch, but was built on the knowledge provided by pre-trained models instead. Fig. 4 describes the transfer learning process used in this study.

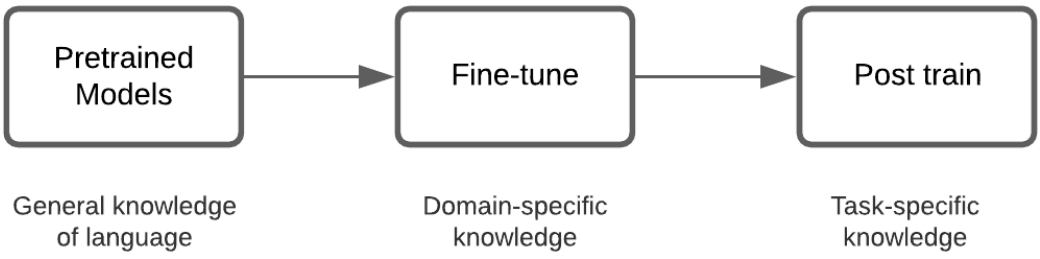

Figure 4: Overview of transfer learning

Pre-trained models are models created previously to solve a similar problem. Those models usually require a huge amount of computational resources, have millions even billions of parameters, and are trained on large, general corpus, such as Wikipedia and

Google News. Pre-trained models offer great general language knowledge, which benefits the target task. In this study, four pre-trained models (BERT [27], DistilBERT [28], XLNET [29], ELECTRA [25]) were used as the starting points.

Pre-trained models require further fine-tuning so they can be used for target tasks. Fine-tuning aims to drop the top layers of the deep learning models, representing the old problem-specific features and to create new layers to inject domain knowledge specific to this study, which could improve the performance of pre-trained models. The data from the consolidated database was used to fine-tune four pre-trianed models in this study.

These four NLP models already possessed general language knowledge from pre-trained corpus, and task- and domain-specific knowledge provided by human insights can further improve the performance. First, these four pre-trained NLP models were fine-tuned using a general causality extraction dataset from the Task 8 [9] in SemEval 2010 to gain task-specific knowledge. Then, they were post-trained [30] in our labeled CA DMV dataset to gain domain- and task-specific knowledge to further improve the performance.

3.3.2 Target Task

During the transfer learning process, the target task was defined with two different approaches to maximize the benefits of transfer learning.

The first approach utilized two same-type pre-trained models with different heads for different purposes - Model One with a token classification head was fine-tuned for cause-and-effect relationship extraction and Model Two with a sequence classification head was fine-tuned for cause category classification. The two models were chained together, so the extracted causes from Model One were fed into Model Two directly to obtain their categories. This approach allowed Model One to benefit from post-training but dropped the sentence context for Model Two.

The second approach only involved one end-to-end model. The tagged labels and categories were combined (Table 8 shows how the combination works) to satisfy the requirements of the end-to-end model. Compared with the first approach, the second approach was more computational efficient and enabled the usage of sentence context for category prediction.

Table 8: How label and category were combined

Words	Label	Category	Combined Label
driver	B-E		B-E
disengagement	I-E		I-E
due	O		O
to	O		O
planning	B-C	2 - planning	B-C-2
discrepancy	I-C	2 - planning	I-C-2

3.4 Data Analysis

Once the best model was identified for cause-and-effect extraction, the final disengagement database was generated, which can be filtered by year, manufacturer, initiator,

location, and cause categories. Then the filtered data were further analyzed using taxonomy and visualization and statistical analysis. With taxonomy, the AVD causes were classified into categories and subcategories. The frequency and distribution of causes in different categories revealed insights, such as "in which stage, the AV system failed to execute tasks most frequently" and "which environmental factors caused AVD most often". Various visualization presented more direct and intuitive information. For example, word cloud gave impression of those hot words used frequently to describe AVD while time-series graph showed how the cases of AVD changed over the past 7 years and demonstrated some patterns which deserved further investigation. On the other hand, statistical analysis provided more quantitative findings about significant relationships between variables with statistical tests.

4 Results

4.1 Cause-Effect Extraction

Following the best practice of transfer learning, the four pre-trained models were all fine-tuned using AdamW [31] as the optimizer with an initial learning rate of 5e−5 and a learning rate scheduler decreasing the learning rate linearly on 15 epochs. To investigate the generalization of the models on new data, 5-fold cross validation was applied. The whole labeled dataset was partitioned evenly into 5 complementary subsets based on the distribution of labels. Each subset was used for testing once while the rest four subsets were used for training. In this manner, the whole dataset was fully utilized and the scores of the models were averaged over subsets to generate the final balanced score.

As Table 9 shows, among the four pre-trained models, ELECTRA achieved the highest weighted F-1 score with relatively low computation resources. Also, post-training significantly improved the performance of the best model at the cost of longer training time. The pre-trained model with fine-tuning and post-training achieved the best performance.

Table 9: Weighted F-1 score and training time for cause-effect extraction

Model	**Weighted F-1**	**Training Time**
BERT + Fine-tuning	0.76	17:39
XLNET + Fine-tuning	0.78	24:52
DistillBERT + Fine-tuning	0.75	10:30
ELECTRA + Fine-tuning	0.82	17:46
ELECTRA + Post training	0.69	12:36
ELECTRA + Fine-tuning + Post training	**0.90**	**28:14**

4.2 End-to-End Token Classification

The training setup for this approach was the same as the first approach. As more specific labels were used for the end-to-end model, the difficulty and complexity for prediction increased a lot. But as Table 10 shows, the pre-trained models still achieved relatively good performance. If Table 9 was compared to Table 10, it is satisfactory to find that the

weighted F-1 score of ELECTRA model with fine-tuning only dropped from 82% to 75% while the complexity of token classification task scaled up 8 times from 7 tags (O, B-C, I-C, B-E, I-E, B-CE, I-CE) to 55 tags (O, B-C-0..., I-C-0..., B-E-0..., I-E-0..., B-CE-0..., I-CE-0...).

Table 10: Weighted F-1 score and training time for end-to-end token classification

Model	**Wighted F-1**	**Training Time**
BERT + Fine-tuning	0.72	17:59
XLNET + Fine-tuning	0.74	24:57
DistillBERT + Fine-tuning	0.71	10:37
ELECTRA + Fine-tuning	**0.75**	**18:12**

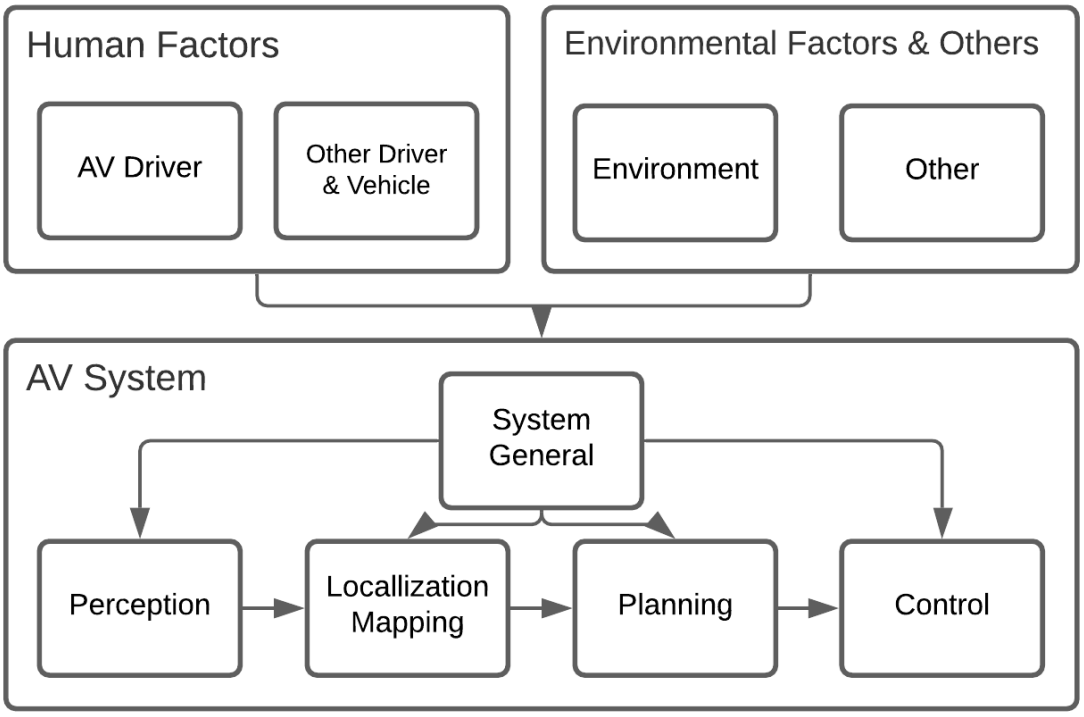

Figure 5: Overview of the taxonomy

4.3 Taxonomy

Schaefer et al. [22] used a meta-analysis method to identify three major factors influencing the trust development in automation, including human factors, system factors, and environment factors. This potentially caused the operators to initiate the disengagement due to a lack of trust [32]. Similarly, we adopted this method to categorize the causes of the disengagement. Furthermore, the failure of different components - perception, planning, control [33] in AV system also led to disengagement. Such failures could be caused

by the environment and the system itself. Thus, a taxonomy (Fig. 5) covering both human driver trust and AV system failures, and environmental factors was generated. This taxonomy served as the basis of the categories and subcategories for causes, which made further analysis more intuitive and efficient. Table 11 shows the number of unique causes in each category.

Table 11: The number of unique causes in different categories

Cause Category	Count
0 - perception	52
1 - localization & mapping	43
2 - planning	52
3 - control	59
4 - AV driver	41
5 - other driver & vehicle	60
6 - environment	33
7 - system general	29
8 - other	8

4.4 Visualization

The pipeline can produce various visualizations to support further analysis, including but not limited to the word cloud of causes under different categories (Fig. 6), the time series of disengagement reported by manufacturers in specific time range (Fig. 7), the multi-series bar chart for causes initiated by the AV system or the human driver, etc.

Figure 6: Sample word cloud for the category environment

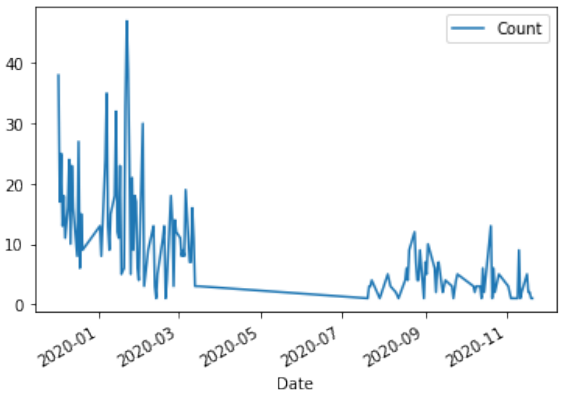

Figure 7: Sample disengagement time series chart for Toyota in 2020

Fig. 8 demonstrated how cases of reported AVD for top 5 manufacturers changed over time. From the figure, it was easy to find that Cruise started AV testing very early but paused at the end of 2018, while other manufacturers continued to test AVs until the early 2020 when Covid-19 happened. After Covid-19 was gradually under control, they resumed the tests but had not reached to their full capacity yet. Some patterns can also be identified from the figure. For instance, those peaks in the chart always happened in the Winter and/or Spring seasons: (1) Cruise - 2015 Winter, 2016 Spring, (2) Lyft -

2019 Spring, (3) Mercedes Benz - 2018 Spring, 2019 Spring, 2019 Winter, 2020 Spring, (4) Nvidia - 2019 Spring, (5) Toyota - 2019 Winter, 2020 Spring. Further researches are needed to investigate whether it was a coincidence or there were reasons for manufacturers to test AVs intensively during these two seasons.

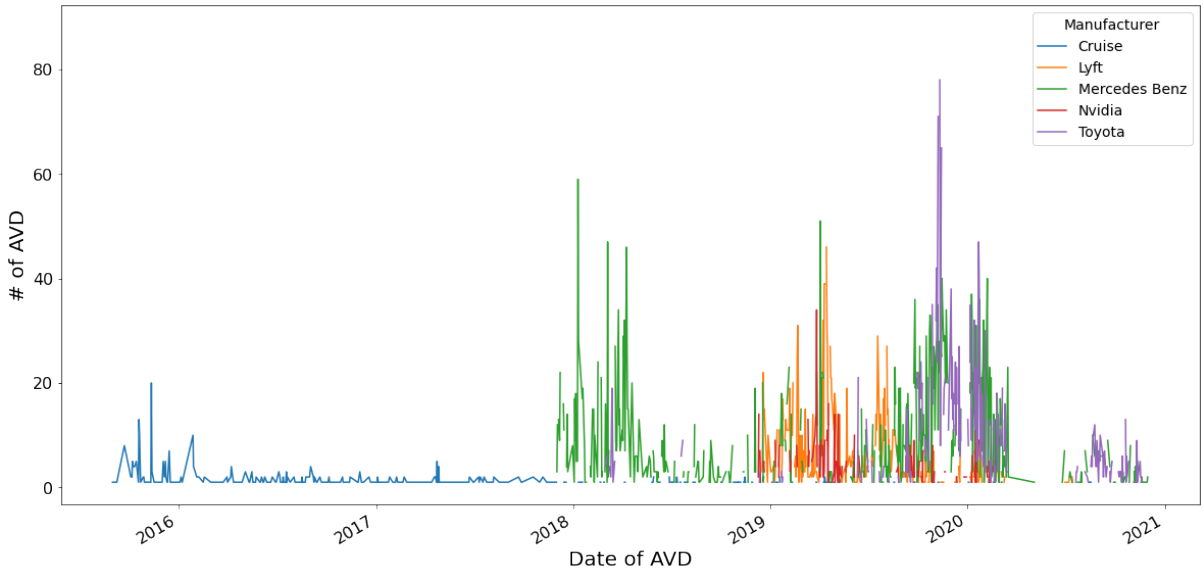

Figure 8: Time series chart for top5 manufacturers having the most AVD counts

4.5 Statistical Analysis

The pipeline can also conduct multiple statistical tests to answer questions such as "Is there a significant relationship between the initiator and the cause category?" or "How did the location of AVD correlated with the cause category?". Table 12 and Table 13 were the contingency tables for initiators between the categories and the subcategories. Table 12 showed that for AVD initiated by AV systems and test drivers, most causes came from the AV system itself. Table 13 provided a more detailed insight suggesting that for AVD initiated by the AV system, the planning stage was the most unreliable stage in the AV system, while for AVD initiated by the test driver, the majority were caused by either the control stage of the AV system or the dissatisfaction felt by human driver. In addition, Chi-Square tests for independence were conducted on the two contingency tables. There was a significant relationship between initiator of AVD and main category, χ^2 (2, N = 9511) = 571.53, p <0.001. And the frequency of causes in the subcategories differed by the initiators as well, χ^2 (8, N = 9511) = 1726.13, p <0.001.

Table 12: The contingency table for initiator and cause category

Initiator / **Cause Category**	AV System	Test Driver
AV System	1703	5493
Human Factors	1	1871
Environmental Factors and Others	46	397

Table 13: The contingency table for initiator and sub cause category

Initiator / **Sub Cause Category**	AV System	Test Driver
0 - perception	322	998
1 - localization & mapping	106	221
2 - planning	775	1423
3 - control	71	2291
4 - AV driver	0	1534
5 - other driver & vehicle	1	337
6 - environment	38	185
7 - system general	429	560
8 - other	8	212

5 Discussion

The goal of this study was to create a scalable end-to-end pipeline to analyze past and future disengagement reports submitted to CA DMV by manufacturers. The two research questions of this study were to identify the cause-and-effect relationships from disengagement data and classify causes into categories and subcategories with a taxonomy. The result suggested that the causes of disengagement were identified and classified into three main categories, including human factors, AV system, and enviornmental factors & others, which were further divided into 9 subcategories: Perception, Localization & Mapping, Planning, Control, System General, AV Driver, Other Driver Vehicle, Environment, Other. Further analysis on cause categories, cause frequency, and cause initiators provided valuable insights into current issues existed in the AV system and potential improvement for manufacturers to improve AV safety.

The findings of this study have important implications for AVD and AV design and testing. The analysis of disengagement initiators suggests that more than 80% of the disengagement were initiated by test drivers, who either felt uncomfortable about the maneuver of the AVs or made precautionary takeover because of insufficient trust. Therefore, additional researches on human trust towards AV and human comfort levels are important to further investigate and explain the manual takeover, which will be a solid step to solve unnecessary disengagement and lay the foundation for the future full automation. The results of various causes related to the AV system also suggest that discrepancy happened in perception, localization & mapping, planning and control was the primary reason that led to the failure of executing certain tasks by the AV system. The majority of prior studies focusing on exploration of disengagement causes using taxonomy had successfully concluded the categories [12], but they didn't identify the specific causes, such as "a wrong speed control command" or "software module generated a wrong path and froze" that were valuable to manufacturers in term of improving the AV system design and testing.

The methods used in this study also have implications for future AVD analysis. As Bimbraw et al. [34] concluded "Most cars are expected to be fully autonomous by 2035", before that actually happens, AV testing still requires the presence of human operators which will generate a large amount of disengagement reports. However, majority of

previous studies heavily relied on the manual work and their findings can not be applied to incoming disengagement reports seamlessly. The weighted F-1 score of this study suggested that with transfer learning, pre-trained models with millions of parameters can be fine-tuned and post-trained to learn the domain knowledge of AVD and the task knowledge of specific analytical tasks, thus achieving satisfying performance close to human workers with much less cost and time. Furthermore, because of the scale of the future disengagement reports, the end-to-end pipeline approach used in this study is both efficient and necessary for the large-scale analysis of AVD that other researchers may be interested in. As some fields of the disengagement reports like vehicle identification number (VIN) and capability of operating without a driver, were not fully utilized in this study, there is also a need to expand and enrich the features in data analysis for the pipeline. Other researcher can build their own version of pipelines based on the one presented in this study while alter the stages and functionalities of pipelines to satisfy their own needs.

Compared with previous research, this study utilized the most complete disengagement reports from 2014 to 2020. Besides, it followed the best practice of NLP deep transfer learning with a novel post training to achieve better performance. Some limitations also existed in this study. Even the performance of the best model was good enough, it is still not as good as human. Also, this study required additional computational resource, especially GPU to train models and build the pipeline.

So far, only three workers provided insight towards cause-and-effect relationships and cause categories, which may cause bias. More workers are needed to make the ground truth labels more reliable. Also it's beneficial to other researchers interested in the disengagement database by adding more statistic analysis features and more customizable functions to the pipeline.

6 Conclusion

This study created a scalable end-to-end pipeline based on NLP deep transfer learning, which can not only collect, process raw data to generate a consolidated database, but also extract and analyze the causes in AVD from the CA DMV dataset. The best model used in the pipeline was the product of best practice of transfer learning followed by post training. A taxonomy covering both human operator trust, AV system safety, and environmental factors was introduced to better understand the causes.

References

[1] SAE. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. SAE International in United States, J3016–201806, June 2018.

[2] Jackie Ayoub, Feng Zhou, Shan Bao, and X Jessie Yang. From manual driving to automated driving: A review of 10 years of AutoUI. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 70–90, 2019.

[3] Na Du, Feng Zhou, Elizabeth M Pulver, Dawn M Tilbury, Lionel P Robert, Anuj K Pradhan, and X Jessie Yang. Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation Research Part C: Emerging Technologies*, 112:78–87, 2020.

[4] Feng Zhou, X Jessie Yang, and Xin Zhang. Takeover transition in autonomous vehicles: A YouTube study. *International Journal of Human–Computer Interaction*, pages 1–12, 2019.

[5] Na Du, Feng Zhou, Elizabeth M Pulver, Dawn M Tilbury, Lionel P Robert, Anuj K Pradhan, and X Jessie Yang. Predicting driver takeover performance in conditionally automated driving. *Accident Analysis & Prevention*, 148:105748, 2020.

[6] Feng Zhou, X Jessie Yang, and Joost de Winter. Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. *arXiv preprint arXiv:2103.14792*, 2021.

[7] State of California Department of Motor Vehicles. Adopted regulatory text. `https://www.dmv.ca.gov/portal/file/adopted-regulatory-text-pdf/`, . Accessed: 2021-04-12.

[8] State of California Department of Motor Vehicles. Autonomous vehicles. `https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/`, . Accessed: 2021-04-12.

[9] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals, 2019.

[10] Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219, Jan 2021. ISSN 0925-2312. doi: 10.1016/j.neucom.2020.08.078. URL `http://dx.doi.org/10.1016/j.neucom.2020.08.078`.

[11] Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1007. URL `https://www.aclweb.org/anthology/W16-1007`.

[12] Francesca Favarò, Sky Eurich, and Nazanin Nader. Autonomous vehicles' disengagements: Trends, triggers, and regulatory limitations. *Accident Analysis Prevention*, 110:136–148, 2018. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2017.11.001. URL https://www.sciencedirect.com/science/article/pii/S0001457517303822.

[13] Alexandra M. Boggs, Ramin Arvin, and Asad J. Khattak. Exploring the who, what, when, where, and why of automated vehicle disengagements. *Accident Analysis Prevention*, 136:105406, 2020. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2019.105406. URL https://www.sciencedirect.com/science/article/pii/S000145751931019X.

[14] Song Wang and Zhixia Li. Exploring causes and effects of automated vehicle disengagement using statistical modeling and classification tree based on field test data. *Accident Analysis Prevention*, 129:44–54, 2019. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2019.04.015. URL https://www.sciencedirect.com/science/article/pii/S0001457519300016.

[15] Hananeh Alambeigi, Anthony D. McDonald, and Srinivas R. Tankasala. Crash themes in automated vehicles: A topic modeling analysis of the california department of motor vehicles automated vehicle crash database, 2020.

[16] Vindhya Venkatraman, John Lee, and Chris Schwarz. Steer or brake? modeling drivers' collision avoidance behavior using perceptual cues. *Transportation Research Record Journal of the Transportation Research Board*, 16-6657, 01 2016. doi: 10.3141/2602-12.

[17] Gustav Markkula, Richard Romano, Ruth Madigan, Charles W Fox, Oscar T Giles, and Natasha Merat. Models of human decision-making as tools for estimating and optimizing impacts of vehicle automation. *Transportation research record*, 2672(37): 153–163, 2018.

[18] S. S. Banerjee, S. Jha, J. Cyriac, Z. T. Kalbarczyk, and R. K. Iyer. Hands off the wheel in autonomous vehicles?: A systems perspective on over a million miles of field data. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 586–597, 2018. doi: 10.1109/DSN.2018.00066.

[19] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[20] Ray Smith. An overview of the tesseract ocr engine. In *Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633, 2007.

[21] Adrian Rosebrock. Ocr a document, form, or invoice with tesseract, opencv, and python. https://www.pyimagesearch.com/2020/09/07/ocr-a-document-form-or-invoice-with-tesseract-opencv-and-python/. Accessed: 2021-04-12.

[22] Kristin Schaefer, Jessie Chen, James Szalma, and Peter Hancock. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58, 03 2016. doi: 10.1177/0018720816634228.

[23] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. 2018. URL `https://arxiv.org/abs/1808.06080`.

[24] Alexandra M. Boggs, Behram Wali, and Asad J. Khattak. Exploratory analysis of automated vehicle crashes in california: A text analytics hierarchical bayesian heterogeneity-based approach. *Accident Analysis Prevention*, 135:105354, 2020. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2019.105354. URL `https://www.sciencedirect.com/science/article/pii/S0001457519308735`.

[25] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

[26] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[29] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[30] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. Bert post-training for review reading comprehension and aspect-based sentiment analysis, 2019.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[32] Jackie Ayoub, X Jessie Yang, and Feng Zhou. Modeling dispositional and initial learned trust in automated vehicles with predictability and explainability. *Transportation research part F: traffic psychology and behaviour*, 77:102–116, 2021.

[33] Scott Drew Pendleton, Hans Andersen, Xinxin Du, Xiaotong Shen, Malika Meghjani, You Hong Eng, Daniela Rus, and Marcelo H. Ang. Perception, planning, control, and coordination for autonomous vehicles. *Machines*, 5(1), 2017. ISSN 2075-1702. doi: 10.3390/machines5010006. URL `https://www.mdpi.com/2075-1702/5/1/6`.

[34] Keshav Bimbraw. Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology. In *2015 12th international conference on informatics in control, automation and robotics (ICINCO)*, volume 1, pages 191–198. IEEE, 2015.