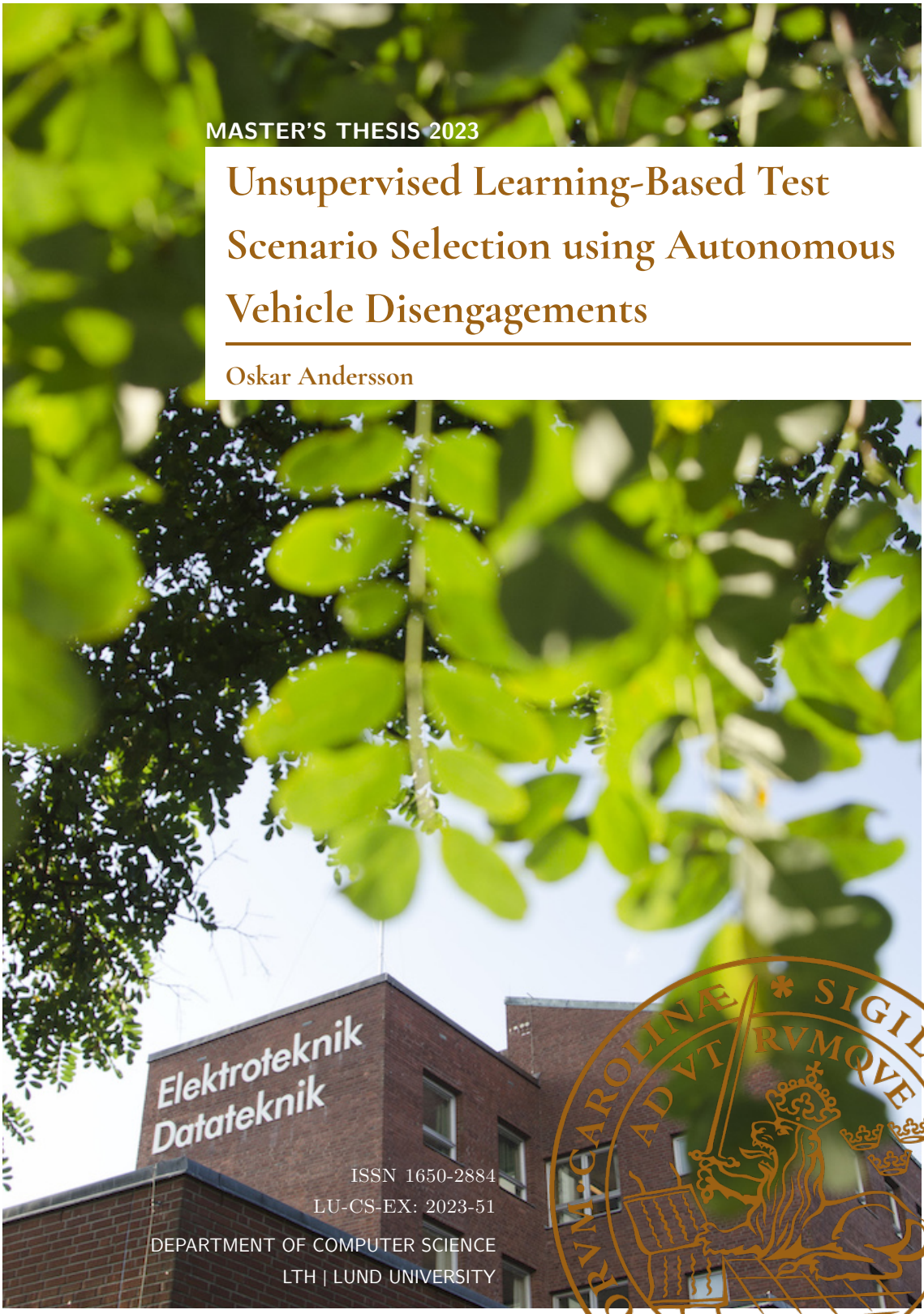

MASTER'S THESIS 2023

Unsupervised Learning-Based Test Scenario Selection using Autonomous Vehicle Disengagements

Oskar Andersson

ISSN 1650-2884
LU-CS-EX: 2023-51

DEPARTMENT OF COMPUTER SCIENCE
LTH | LUND UNIVERSITY

EXAMENSARBETE
Datavetenskap

LU-CS-EX: 2023-51

Unsupervised Learning-Based Test Scenario Selection using Autonomous Vehicle Disengagements

Testscenariourval från frånkopplingsrapporter för förarlösa bilar genom oövervakad maskininlärning

Oskar Andersson

Unsupervised Learning-Based Test Scenario Selection using Autonomous Vehicle Disengagements

Oskar Andersson
os8675an-s@student.lu.se

December 7, 2023

Master's thesis work carried out at
the Department of Computer Science, Lund University.

Supervisor: Qunying Song, qunying.song@cs.lth.se
Examiner: Emelie Engström, emelie.engstrom@cs.lth.se

Abstract

Autonomous vehicles are becoming a subject of testing on public roads to ensure their safety before making them publicly available. With the complexity of their operation, new testing routines and standards needs to be implemented and evaluated to ensure safe operation. Many research papers on the subject apply a scenario-based testing methodology with the principle of finding representative test scenarios and ensuring that the system performs appropriately for these. Currently the field is faced with challenges in articulating test scenarios and making sure that they capture all possible scenarios. I use a scenario-based testing approach combined with unsupervised learning to find representative scenarios automatically from realistic autonomous vehicle disengagements. The resulting clusters are evaluated to determine what form of vectorization and embedding of textual entries leads to the most accurate results. The results from clustering were that the methodology was able to produce clusters with high performance in regards to three common clustering metrics for a data set of 184 disengagement entries. The evaluation of the actual scenarios that the methodology was able to cover did however not indicate that the methodology achieved a high level of accuracy, with the highest percentage achieved being approximately 41% for KMeans clustering with 19 clusters and approximately 53% coverage with 35 clusters using DBScan. In conclusion, the report reveals the methodology as a feasible way to mine test scenarios. However, the lack of large data sets of disengagements makes the tool hard to conclusively evaluate and the similarity comparison between disengagement scenarios is hindered by the lack of embeddings specialized in semantics in the field of autonomous vehicles.

Keywords

Test selection, Unsupervised learning, Autonomous driving systems, Scenario-based testing, Autonomous Vehicle Disengagement

Acknowledgements

I would like to express my sincere gratitude to all those that have contributed to the completion of this degree project. This thesis would not exist if it weren't for their invaluable support and guidance.

First and foremost, I extend my greatest appreciation to my supervisor Qunying Song. For taking on the project and providing me with a constant flow of constructive feedback and expertise in the field. Though the research goals at times were hard to progress with, the guidance from Qunying Song helped me keep focus on the goals of the project.

Secondly I would also like to give a thank you to my examiner Emelie Engström. For taking on the role as examiner and in that role helping this thesis reach a higher quality.

Contents

1

Introduction

Autonomous vehicles are improving and becoming ever more present in the public eye [1]. The technology is thought to have many possible benefits such as achieving more safe roads considering the presence of human error in 93% of current accidents. There are also many other benefits such as better infrastructure planning as population increases, leading to more vehicles on the road [2], [3].

Autonomous vehicles are a subgroup of autonomous systems which are systems capable of operating on their own without human interaction and other forms of autonomous systems are currently present in the everyday life [4]. Though autonomous systems are not foreign, there are concerns relating to the evolution of autonomous vehicles, where the safety related concerns are one of the most prominent concerns holding autonomous vehicles from the public roads [5].

Concerns regarding the safety of the autonomous vehicles and the accuracy of their decisions are present in the debate [6]. Multiple accidents where autonomous vehicles have been involved have been reported, where some have had fatal outcome revealing that the systems are not flawless and these flaws are important to address. In order to reap the benefits of autonomous vehicles, there needs to be a very low risk of operation which makes testing a central problem to further expansion of autonomous vehicles [7].

However, the testing of autonomous vehicles is not a simple task. The way context plays a role in their behavior and the infinitely large possible scenarios that the vehicle can be involved in requires new forms of testing methodologies [4],[8].

The large and varying number of scenarios an autonomous vehicle needs to take into consideration makes fully exhaustive testing an impractical option as the identification of all possible scenarios is infeasible due to limited testing resources. The process of performing on the road real-world testing of autonomous vehicles have been estimated to require upwards to 11 billion miles in order to reach a 95% confidence that autonomous vehicles are 20% more safe than human operated vehicles [9], [10], [11]. Instead of using a mileage-based testing of vehicles, many suggests using a scenario-based testing approach [11], [12], [13]. Such an approach is based on finding representative test scenarios that can test the system and make sure that the scenario representation set covers the possible scenarios that an autonomous

vehicle can be part of while in operation. This data is essential to build a scenario database to perform testing based upon [9]. This process is currently adopted and developed to address the problems of finding and formalizing the scenarios that are relevant to increase confidence in testing and evaluate test coverage [4].

As an alternative to achieving the coverage purely based on real-world testing, many autonomous vehicle manufacturers use simulation-based testing, where many different driving scenarios can be generated and tested to see the autonomous vehicle response for a reduced cost [14]. There are however concerns over the quality and coverage of scenarios in simulation testing against real-world testing and therefore real-world driving is still a part of the testing to bridge the gap in simulation testing [15].

In order to test the autonomous vehicles on real-world scenarios, the state of California has passed legislation which allows manufacturers to test autonomous vehicles on their public roads [7]. Some guidelines are however necessary to follow for the actors wanting to perform testing on public roads and one part of the requirements is that the manufacturers must submit reports annually of the autonomous vehicle disengagements. The reports include information such as when and where a disengagement occurred as well as a summary of the situation. The reports are publicly available on the California Department of Motor Vehicles (California DMV) homepage [16]. Such data is expected to improve the generation and selection of relevant test cases [17] and enhance safety of autonomous vehicles [18].

However, finding representative scenarios from realistic AV disengagements remains an open challenge, and this project aims to address this by taking an unsupervised approach to extracting representative disengagement for test scenario selection and creation for autonomous vehicles as well as providing measurements to evaluate how such a model performs.

The main goal of the thesis is to achieve an efficient and effective way of selecting test scenarios from real-world driving disengagement data for autonomous vehicles from California DMV based on unsupervised learning.

The paper is structured into 7 sections where the first section is the introduction. Section 2 named "Concepts and Related work presents works related to the thesis as well as concepts applied in the thesis. Section 3 named "Research approach" presents how I aim to fulfill the main goal of the thesis. Section 4 named "Scenario Selection Methodology" presents the approach taken to achieve scenario selection according to the thesis main goal. Section 5 named "Evaluation" shows the evaluation of the methodology that has been devised. Section 6 named "Discussion", the research questions are discussed using the results achieved in section 5. Lastly, section 7 shows the conclusions that are arrived at using the discussion.

2

Concepts and Related Work

2.1 Concepts and terms

2.1.1 Testing of autonomous vehicles

2.1.1.a Industrial standards

The level of AV autonomy is commonly described by 5 levels of autonomy according to the Society Of Automotive (SAE) standard SAE-J3016 [19]. This ranges from no automation to fully autonomous vehicles that can drive under all conditions without human intervention.

Some validation standards related to testing autonomous vehicles exist such as ISO 26262:2018 [20]. This standard mainly focuses on the security and safety of electronic hardware under the presumption that a responsible driver is in control of the vehicle [21]. This standard places emphasis on ensuring that the electronics do not break during usage.

There is another standard for ensuring that the components of vehicles work as intended without failure in the form of ISO 21448:2022 [22]. This standard places the focus more on functionality, where the standard requires that the electronics can behave correctly even in the case of unexpected scenarios.

These standards does however have a common issue which limits their usability in autonomous vehicles as they mainly state general requirements and limited scenarios for testing, which is insufficient to implement and ensure the safety of autonomous vehicles [10], [23], [24]. The common challenge that is identified in the standard is the way testing using current standards requires the presence of a safety driver to handle the vehicle in the cases where the integrated features of the vehicle fails. There are also concerns regarding the level of unpredictability present in machine learning technology and how the standards should be reformulated to implement a standard validation process for autonomous vehicles. Koopman and Wagner also describes the infeasibility of performing complete system-level testing as a way of validation due to the time requirement which is estimated to be many billion hours [23]. The lack of internationally accepted requirements on testing cur-

rently poses problems with defining what type of testing should be performed for autonomous vehicles according to Khastgir et al. [13].

2.1.1.b Challenges

Due to the problems regarding standards present in electrical equipment for vehicles, challenges exist for testing autonomous vehicles. Without a comprehensive standard for testing of autonomous vehicles, much effort is put into finding effective ways of test functionality and safety of autonomous vehicles [25]. There are other non-technical factors at play such as the ethics of decision-making, this thesis will not cover such factors as they are out of the scope of this thesis [23].

Some characteristics that make the validation of autonomous vehicle software different from other forms of software validation is the complexity of the software involved in the decision making, which relies on machine learning and other statistical algorithms [2], [23]. The result is that the tests conducted can have different results even though the same type of information is being inputted. Considering that many of the software present in AVs is based on ML models, the systems often have to be seen as black-box systems [26]. These core features are seen by studies also as the main reasons for why validation is central to AVs as there is a possibility that the models that are trained have not been trained on all scenarios [27]. The risk of so called "black-swans" in the system, which may be rare, contributes to the vast number of mileage necessary to cover in on-road testing [13]. Therefore, it is possible that the training did not cover all possible scenarios and such scenarios needs to be revealed to improve performance and safety [27].

2.1.1.c Current practices

Considering the challenges described in Section 2.1.1.b, only using a mileage-based testing approach is inadequate for autonomous vehicles [12]. Current approaches are developing towards finding the most critical scenarios instead of just performing on-road testing [8], [13]. And also finding scenarios that can be representative of a larger set of concrete scenarios [11] to handle the problem of having infinite amounts of scenarios that needs to be tested individually. The critical scenarios are seen as the most important scenarios to validate. This can be based on the severity of consequences incorrect decisions in the scenario can result in. The benefits of using real-world validation is that the AVs actual performance can be assessed [11]. However, such testing is expensive, which makes simulation testing a significant approach in AV testing [25], [26], [27], [28].

Xu et al. shows that there is a challenge to find and formulate ground truths of expected outcome for a certain test input set due to the vast input space for a complete autonomous vehicle system [29]. A conventional approach could be performing domain-categorization of the input and see if the system response is as expected for each domain [30]. With the difficulty of determining desired outputs

based on singular inputs to the system, and the complexity of surrounding information to make a decision [31], [4], many studies advocate a scenario-based approach to testing autonomous vehicles[4], [30], [32], [33], [34].

The lack of unified standard for validating AVs [25] makes the decision on a test ending criterion to achieve desired confidence of functionality difficult. Hauer et al. describes the possible way to determine a test ending criterion for scenario-based testing is to identify realistic scenarios and then to have them tested for expected functionality [30]. According to Lou et al., identification of unexpected driving scenarios, determining if all similar scenarios result in the same autonomous vehicle behavior and finding efficient ways to perform validation are the most crucial improvements needed in the industry [35]. Wagner and Koopman also state that the edge cases are more important to find and investigate rather than performing confirmatory testing of functionality that has already been validated to some extent [36].

2.1.2 Scenario-based testing for Autonomous Vehicles

Scenario-based testing is commonly used in autonomous vehicle testing [8]. With many non-scenario based testing approaches the problem arises of identifying relevant parameters and combinations of parameters to test [31], [37]. Often, the importance in AV validation is to determine if the system behaves appropriately when presented with a certain situation rather than seeing the behavior for certain possible values of the system parameters. Scenario-based testing can also be a possible solution to the problem of validating using only distance-based on the road testing, as the goal is instead to cover the scenarios which may occur and validate that the AVs respond appropriately to similar scenarios [12].

2.1.2.a Definition of a scenario

Extensive studies within the field of scenario-based testing uses the definitions of scenarios defined by Menzel et al, [25], [34], [37], [38]. The authors describe scenarios in three levels of abstraction with logical, concrete and functional scenarios and their application in different stages of development.

- *Functional scenarios* are the highest form of abstraction presented in Menzel et al., where scenarios are usually described verbally or in text. Menzel et al. considers this level as an appropriate way to describe scenarios during the early stages of development according to the ISO 26262 as it makes it easy to interpret the intention of the scenario and to design new complementary scenarios for testing.

- *Logical scenarios* are the second highest form of abstraction and has a more detailed description of the scenario than functional scenarios. With this level, the parameters are given possible ranges and distributions which can make

up a scenario. For example, a functional scenario state that a lane is wide and the logical scenario should define the measurement and boundaries for a road lane width. This abstraction level is seen as a good level to set the requirements for the scenario as the necessary parameters and their values have been determined.

- *Concrete scenarios* are the lowest form of abstraction. They are specific situations where the parameters determined in the logical scenarios are assigned one concrete value. Menzel et al. sees the role of concrete scenarios as a way to design test cases as they are based on testing the system response for different input parameters and parameter values.

2.1.2.b Scenario Parameterization

The scenarios can be described according to the definitions presented in 2.1.2.a, but in order to do so with a systematic approach the scenarios can be parameterized. Goss *et al.* [39] and Bach *et al.* [40] have authored two studies that define logical scenarios by determining parameters present in the scenarios and representing them by the values of those parameters and the intentions of the involved actors. They use the parameters and their values as atomic portions of the scenarios in order to represent the situations in a cohesive way and also to form possible scenarios by combining new parameter and parameter values to find scenarios that could possibly be relevant to test as well. Parameterization as a way of test scenario generation is also seen as a viable approach by Gelder and Paardekooper [11]. An example of the parameterization can be seen in table 2.1. When the parameters and their values have been determined, Wotawa shows the possibility of using T-way combinatorial testing to expand the possible scenarios that can be covered [41]. The T-way testing requires all combinations of all parameter values to be tested in order to fully test the system for all possible inputs and their interplay with each other.

Table 2.1 An example of parameterization for a disengagement entry in the California DMV database.

Description	Parameterized representation
Car stopped in middle of intersection and did not proceed after pedestrian was finished crossing crosswalk	Intersection, Crosswalk, Pedestrian, Cross street, legal

2.1.2.c Localizing scenarios to test

In order to perform scenario-based testing of autonomous vehicles, the tester must have a scenario database containing required scenarios to reach the desired confidence in functionality [9]. It is not only the process of formalizing and describing

the scenarios that is necessary to perform scenario-based testing, but also to determine which scenarios are relevant to test. One approach of deriving test scenarios is to select and recreate real-world driving incident reports or studies [9], [39]. Weber *et al.* considers accident data as a better way to derive scenarios to test rather than general driving data as the critical scenarios are more likely to be those that may result in accidents [37].

Esenturk *et al.* studied mining possible scenarios for autonomous vehicle validation [3]. Specifically, they have given high level abstractions of the parameters present in different accidents, such as labeling scenarios where wind was present as "windy". They rely on the testers to determine the values needed to be tested based on these high level descriptions, where the testers design rules for what values constitutes a certain categorical value such as windy.

2.2 Related work

2.2.1 Data-driven scenario generation

Some work has previously been done on using real-world driving data to find possible testing scenarios. Langner *et al.* use a clustering approach to present logical scenarios derived from concrete scenarios collected during real-world testing of autonomous vehicles [33]. They aim to build a logical scenario catalog which can be used both to assess the level of coverage that have been achieved in testing based on the number of scenarios. They also claim it as a tool to test autonomous vehicles with. The problem is similar to the one that is posed in this thesis regarding extracting scenarios from real world driving, but their evaluation process is centered around manual inspection of clusters. Which our thesis propose an alternative to. This is described in Section 4.

The process of clustering is also used in a UK-based study, which focus on accident data from the United Kingdom to identify patterns which may result in traffic accidents, and are used to generate scenarios for autonomous vehicle testing [42]. The study does however not handle the task of determining what coverage can be achieved by their data mining approach, which is explored and reported in this thesis.

2.2.2 California DMV dataset

The CA DMV dataset [16] has been the subject of many previous research projects within the context of testing autonomous vehicles. Zhang *et al.* used this data to reveal which factors most commonly lead to disengagements [43]. They used natural language processing in order to produce a supervised learning model to find cause-effect relationships between the parameters present in disengagement and disengagements. They found the model as a significant benefit to the field as the manual processing of disengagement entries is time-consuming with the autonomous vehi-

cle technology advancement and expansion in the future. This study does however mainly focus on categorizing the different scenarios and does not place emphasis on the testing of systems using categorization of disengagement causes.

Numerous studies have also worked with investigating the disengagement reports manually to determine possible topics to sort the disengagements in categories for gaining insight in security problems of AVs [17], [44], [45],[46]. These studies provide insight into the problem with the formatting of the disengagement entries. They also observed a problem with how some disengagement entries also lack information to fully describe the cause of disengagements. One article also analyzed the data in order to give recommendations to governments and manufacturers for what kind of types of disengagement the systems needs to be improved for in order to increase level of autonomy in AVs [18]. Such studies provide a lot of insight in what one can expect in the reports, but are based on the assumption of that categories are present in the data set. In the field of testing, the desire is to have an unbiased approach to finding possible unexpected faults and by performing manual grouping of scenarios a possible problem of biases arises that unsupervised learning methodologies may be able to mitigate.

2.2.3 Unsupervised learning classification

The task of classifying natural language entries have previously been investigated in other studies. In a study by Lu *et al.*, legal documents are clustered to reveal common topics that are presented in the documents [47]. They classify the documents using soft clustering to allow multiple topics to be assigned to a certain item. The results shows the approach can accurately partition documents into topics. One difference from the application in the autonomous vehicle field is however that the approach was based on a very large dataset, where the features of documents could be assessed for millions of documents rather than a couple of thousand disengagements present in the California DMV dataset [16].

Zhang *et al.* also have presented an approach of finding similar question entries in question and answer archives [48]. Their goal is to find similar questions to supply answers that to other question entries that may be formulated differently semantically, but have the same meaning. This study propose a approach for paraphrasing entries in order to compare the meaning of sentences rather than being phrased in the same ways. The approach is a possible way to achieve less interference of grammatical rules in the similarity comparisons of words. In contrast, we use the simpler approach of lemmatization to achieve comparable results with less complexity.

3

Research approach

The main goal of the thesis is to determine if we can efficiently select test scenarios from real-world driving data using unsupervised learning. My main contributions is implementing a methodology based on unsupervised learning and evaluating its performance. The results of the evaluation will be used to determine its viability in the field. Some aspects of this process is especially important and needs specific attention. Therefore i have devised a set of research questions that will go deeper into the core principles of devising a scenario selection methodology. The research questions are the following:

- RQ1: How can the similarity between different disengagement scenarios be determined?
- RQ2: How can representative disengagement scenarios be selected from the dataset using clustering?
- RQ3: How can test scenarios be formulated to be presented in a standardized way for all reported scenarios?
- RQ4: What coverage of the dataset can be achieved by selecting representative disengagement scenarios?

To group the scenarios by similarity, we need to be able to compare entries. RQ1 is therefore a central topic of investigation to reach the goal. We will investigate this research question by seeing how well common text preprocessing (presented in section 4.3) and embeddings (presented in 4.4) is able to represent disengagements. We also evaluate the results empirically by seeing how large coverage is achievable by using the methodology and how well we capture relations between the disengagements using the embeddings.

We must also determine how we group the data in a way that maximizes the similarity within groups and lead to good test representatives. This makes RQ2 a relevant question to investigate. By using two different clustering algorithms, the

similarities of disengagements will be used to form the best groups for cluster selection. For K-means, this is implemented in section 4.5.1 and evaluated using the clustering coherency metrics defined in 4.6 and visualizations in 5.2. For DBSCAN, the clustering is implemented in section 4.5.2 and evaluated by visualizations in 5.3.

The disengagement descriptions are initially formulated in natural language. However, to evaluate how high test coverage is achieved independently of the clustering evaluation we need to standardize their formulation. Therefore, we must investigate RQ3. In 4.8.1, the process of parameterization is implemented as a possible standardization approach and the approach's efficiency is used in evaluating the coverage in sections 5.2 and 5.3.

Lastly, one central question to investigate is how high coverage we can achieve with the representative scenarios selected using the methodology. In order to show that the usage of unsupervised learning is worth implementing the field, we must investigate RQ4. This research question will be investigated by evaluating the coverage of parameters that can be achieved by the representatives (as described in 4.8.2). The empirical results of the evaluation will be presented in sections 5.2 and 5.3.

4

Scenario Selection Methodology

The main structure of the scenario selection methodology is presented in figure 4.1. Initially, the data is loaded and filtered to exclude uninformative and duplicated data as presented in section 4.1 and 4.2. Then, the data is preprocessed using common text-preprocessing methodologies as described in 4.3. The preprocessed data is then vectorized to represent the data numerically as described in section 4.4. Considering the high dimensionality of the data, dimensionality reduction is also applied, which is also described in section 4.4. Clustering is then performed using the K-means algorithm and DBSCAN as described in section 4.5. The clustering results is also visualized and evaluated as described in 4.6 and 4.7. Finally the scenarios are parameterized and evaluated using the cluster representatives using the methodologies described in 4.8.

The scenario selection methodology is based on a combination of some pre-existing approaches and some unique steps that I have devised. The process of storing the data and filtering the dataset (presented in 4.1 and 4.2) was devised by me. The preprocessing is based on pre-existing methodology, where my contribution is that I identified the need for them to improve the ability to compare similarities between disengagement scenarios. The process of embedding the data is based on previous studies relating to comparing natural language entries. My contribution to the step is the creation of new embeddings based purely on the field of autonomous vehicles. The usage of the K-means algorithm and the DBSCAN algorithm is based on existing implementations of the algorithms. My contribution for these is evaluating them to identify if density-based clustering or distance-based clustering is more appropriate algorithms for the clustering. One large contribution that is made by me is in the evaluation of the clustering and coverage. No pre-existing work has been used to devise the parameter coverage metrics, which enables the evaluation of coverage independently of the evaluation of cluster quality. For a more detailed reference to the studies that have been used for parts of the methodology, I refer the reader to the specific section.

The connection with the research questions is the following. RQ1 primarily relates to the preprocessing and vectorization stage. The process of restructuring the natural language entries and representing them numerically is performed in order to make similarity comparison between disengagement scenarios possible. RQ2 primarily relates to the clustering section of the scenario selection methodology. The clustering efficiency is evaluated using the visual representations, the silhouette score, CH index and DBI. By combining the results of the clustering evaluation with the achieved coverage of test parameters in the dataset, it determines if the clustering approach generates suitable cluster representatives. RQ3 primarily relates to the the scenario selection methodology's ability to represent the disengagement scenarios in a standardized way. This is addressed in the process of parameterizing the disengagement scenarios. RQ4 primarily relates to the step of evaluating test suite coverage. It is this step that makes it possible to determine what coverage is achievable with the scenario selection methodology.

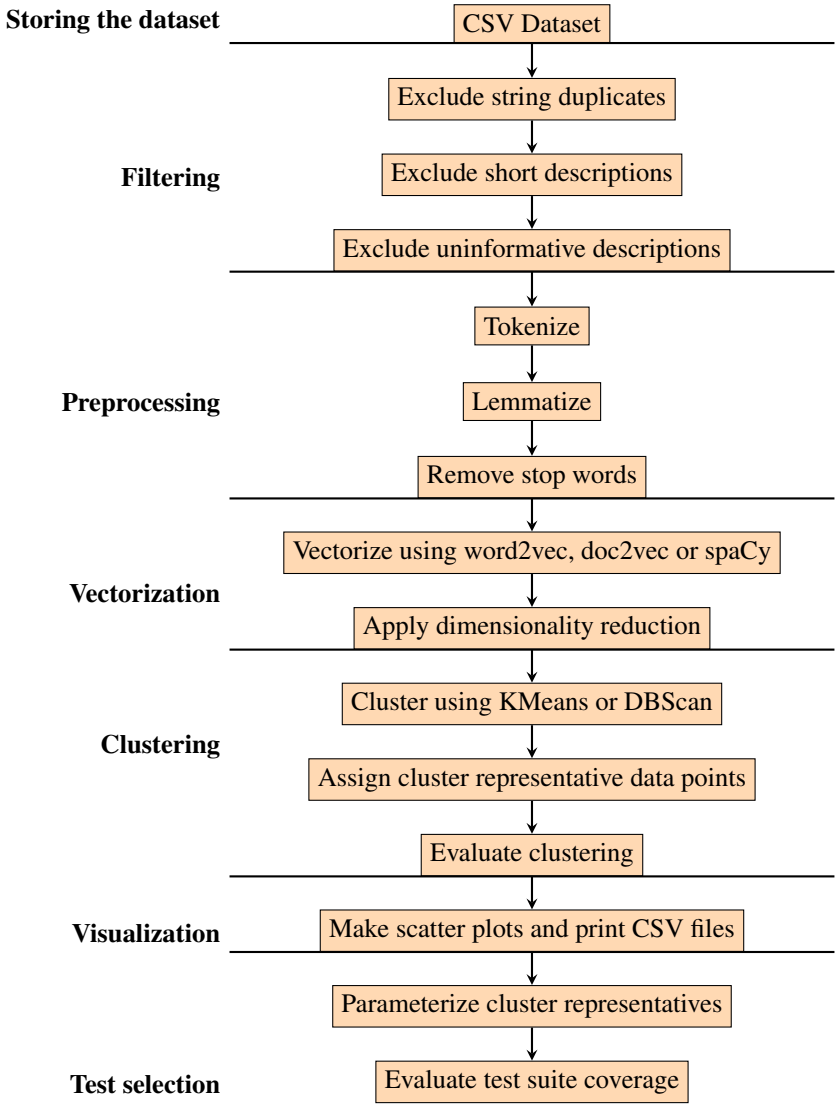

Figure 4.1 The graph shows the processing of the dataset from CA DMV to the point of evaluation.

Manufacturer	**Permit Number**	...	**DESCRIPTION OF FACTS CAUSING DISENGAGE-MENT**
AImotive Inc.	AVT003	...	Lane change maneuver: risk of lane departure, caused by unstable target lane model
Drive.ai Inc.	AVT013	...	System kickout due to hardware due to road surface conditions
...	...	...	...

Table 4.1 The table shows how the data in the CA DMV files are presented. Columns with information about report date, VIN number, vehicle capability of operating without a driver, driver present or not, disengagement initiator (system or driver) and disengagement location have been omitted for brevity.

4.1 Storing the dataset

The input dataset consists of multiple comma separated value (CSV) files fetched from the CA DMV. An example of the information contained in the CSV files can be seen in table 4.1. Each row in the files represent a disengagement situation where the information regarding the disengagement is provided. The description column specifies the situation and cause of the disengagement in natural language. For this thesis, it is the descriptions that is of interest and the data in other columns are excluded from the dataset. The data structure that the dataset is initially formatted as a list of textual entries in order to enable further filtering of the data.

4.2 Filtering the dataset

The dataset comprises a total of 21,351 entries, combining all disengagement descriptions from 2019 through 2022. Earlier descriptions exists, but were not used in this thesis as they were not CSV formatted. However, a lot of these are duplicates and by removing duplicates based on exact string comparisons, the dataset retains 1176 unique textual descriptions. Furthermore, we filtered out possible non-informative entries by setting a threshold corresponding to approximately the length of the average English sentence of 100 characters (20 words per sentence and 5 characters per word) [49]. The threshold is set so that all entries with less than 100

Classification	Disengagement description
FULLY CLEAR	The car was performing a parking maneuver, when the driver had to take control, because the side wall was within 0.5 meters from the car. This event was caused by the error in object detection by s ultrasonic sensor.
PARTIALLY INFORMATIVE	The AV was approaching a turn too fast. As a result, the driver safely disengaged and resumed manual control.
UNINFORMATIVE	Localization/position discrepancy - a problem was observed in the vehicle's estimate of its position; causes may include the accuracy of the imu, lidar, and gps sensor data, the algorithms used to process that data, or the accuracy of the vehicle's maps

Table 4.2 The table shows an example of how disengagement scenarios are classified based on the clarity of their textual descriptions.

characters is removed. This results in a dataset of 552 textual descriptions of autonomous vehicle disengagements.

Still, some long descriptions are also non-informative. We manually filtered out non-informative entries by assigning disengagement entries the labels "uninformative", "partially uninformative" and "fully clear". The classifications are based on if the scenario is described clearly enough to determine a logical and or concrete scenario from. This requires that the disengagement is described with specific details to indicate which concrete scenarios should be associated with the entry. As an example, table 4.2 shows how certain disengagements would be labeled. The uninformative entry is not unclear but rather unspecific. The partially informative entry has some information regarding the scenario, but still proves unspecific and hard to reproduce in the form of a logical or concrete scenario. The last fully clear entry is specific and clearly described in such a way that the cause of the disengagement can be reproduced and tested upon. Within the dataset, there are 11 entries which are identical to other entries in terms of meaning, but have slight differences in the formatting of their descriptions, such as a misplaced white space character. These are also removed from the dataset. As a result, the dataset consists of 184 fully clear disengagement scenarios.

4.3 Preprocessing of the dataset

In information retrieval from natural language texts one article shows common pre-processing steps for extracting relevant information from natural language [50].

The first common step for pre-processing text in natural language, as described by Mohan, [50], is to split the sentences into individual words in order to make the data more easily manipulated. The process is known as tokenization, where the result is a representation where words are individual objects known as tokens. In this thesis, these tokens are stored in lists, where each list represents a disengagement description.

Initially, Mohan describes that the natural language entries usually have some words that carry little information within them, known as stop words [50]. These are functional words often present for grammatical reasons and some examples are words such as "it", "is" and "the". In order to make the texts comparable these stop words are often disregarded in similarity comparisons and therefore removed in the pre-processing steps of natural language processing. The number of stop words that are discarded varies from the dataset and which approach is applied. A common tool in natural language processing known as spaCy has a default set of 326 stop words [51]. spaCy is used in this thesis for stop word removal.

Following this step, Mohan presents the process of stemming, which is used to further handle grammatical differences in the context that words are used in. Stemming is the process of extracting the base of words which may be defined by different endings depending on the situation they are used in. Stemming uses a set of common endings such as "ing" and "er" and remove these endings from all words in order to make them comparable. This however results in the possibility of different words resulting in the same root words as shown by Pramana *et al.* [52]. They consider the usage of lemmatization, which is a method of getting the root of words with the context. They found the lemmatization as a better tool for comparing similarities of sentences compared to stemming in their evaluation. In this thesis, lemmatization is used to reduce the impact that grammatical differences can have on similarity comparisons.

4.4 Vectorization and embedding of the dataset

The task of comparing similarity of texts written in natural language can be performed in different levels. Less complex methodologies such as a string comparison can be used to determine the similarity of texts or words. Such methodologies however do not capture sentences that are structured differently but still have the same meaning. In order to perform a more thorough comparison of the similarity of natural language entries, the context that words are used in and the interplay of words in sentences needs to be taken into consideration.

In a study by Srinivasa-Desikan [53], the author presents multiple tools to deal

with the task of performing semantical comparisons of texts. The author shows how spaCy [51] can be used as an initial way to represent words and sentences in a way that captures the meaning of the sentences. Srinivasa-Desikan also introduces the word embedding tools Word2Vec, Doc2Vec as alternatives to achieve a numerical representation of words and sentences in order to determine their similarities.

With the text tokenized and preprocessed, the task is to capture the relations of the different words and represent them numerically which is done using word embeddings and vectorization [50]. Common embeddings used to capture the relationships between words in texts are word2vec, spaCy and term frequency - inverse document frequency (TF-IDF) [54], [55], [56].

Word2Vec is a neural network based word embedding tool which can learn relations between words in texts [57], [58]. The tool can be used both to train on texts of the users choice or with pre-existing models. One pre-existing model trained by Google for word2vec is word2vec-google-news-300 [59]. I embedded the disengagements descriptions by representing the sentences by calculating the mean matrices of the word2vec representations of the words that make up the disengagement descriptions. Similarly, Doc2Vec is an extension of the Word2Vec implementation that takes a document level approach of similarity and not only single words [60]. The framework has been used by Tahvili *et al.* to reduce a test case suite by finding and filtering out similar cases with clustering based on semantic similarity between test cases [56].

SpaCy is a framework for natural language processing [51]. The model provides tools to preprocess text and give semantic information of each word within a text based on the usage of the word. Models can be trained on corpuses to compare the similarities of words or sentences based on the usage in the corpuses. The module can also be implemented using pre-trained models where the en_core_web_sm and en_core_web_lg models are commonly used [61]. The embeddings for the data was achieved by representing the data with their similarity matrices.

Term Frequency - Inverse Document Frequency (TF-IDF) is an algorithm used to describe the relevance of certain terms within a set of documents [62]. The algorithm calculates the number of occurrences of a term in a certain text collection or document (term frequency). This is weighed towards the number of documents that the term is not present in (inverse document frequency). The result is a measurement that gives an indication of the importance of certain terms and can be used to compare how similar documents are based on the occurrence of words and if they share important words [63].

In section 4.2 and 4.3 we will evaluate different forms of vectorization for clustering to find the best numerical representation for the disengagement descriptions. The following vectorization approaches will be evaluated:

- word2vec trained on the dataset
- word2vec pre-trained model (word2vec-google-news-300)

- doc2vec trained on the dataset
- spaCy pre-trained model (en_core_web_lg)
- TF-IDF

The embeddings have quite a high dimensionality, where the word2vec google news embedding has a dimension of 300. With the sparseness of the data present in the dataset, multiple studies comment on the possibility of a phenomena known as the Curse of dimensionality [64], [65]. These studies present the curse as a problem with using vectors of high dimensionality in categorization of small datasets, where the usage of too large dimension in relation to the dataset may lead to a higher number of features making the accuracy of the classification worse due to inducing noise rather than valuable information about the associations between the data points.

To mitigate this curse of dimensionality for the dataset, different levels of dimensionality reduction are applied in this thesis to find the optimal number of dimensions when grouping the data points. The dimensionality reduced data that achieves the highest scoring is presented for each measurement in sections 4.2 and 4.3. The data is reduced to dimensions ranging from 2 to 99 dimensions as well as in its original dimensions. We use PCA for dimensionality reduction, which have been proven to perform well with K-means clustering approach [66]. The vectorized data are scaled in order to normalize the vectors so that the values within each vector is between 0 and 1.

4.5 Clustering

With a numerical representation of words and sentences as vectors, it is feasible to determine the similarity of sentences mathematically [63] with metrics such as the cosine metric or the euclidean distance metrics [67], [68]. These metrics are in line with unsupervised learning as they have no labeling for the correct assignment to clusters but rather to find clusters that minimize distances to similar data points. The choice of clustering algorithm is dependant on the input data [3], [69] and in order to validate the choice of clustering approaches used in the thesis, two different algorithms are presented, namely K-means and Density based spatial clustering of applications with noise (DBSCAN).

4.5.1 K-means

The K-means algorithm is one of the most commonly used clustering algorithms according to Sinaga and Yang [70]. It has a requirement on the number of clusters to be configured prior to clustering, which is problem dependant and commonly evaluated to see which number of clusters result in the best clustering for the problem [71].

A description for the process of clustering with the algorithm can be found in Na *et al.* [72]. After the number of clusters desired (K) have been set, the algorithm selects K cluster centers at random. The data will then be grouped with a cluster center based on the minimal euclidean distance to the data point to be clustered. The process is repeated until all data points are assigned to a cluster. The initial assignment of the data points to clusters is used to calculate a new centroid for each cluster based on the mean position of all the data points within the cluster. The assignment procedure is repeated with the new cluster centers until there is no change of cluster centroids.

4.5.2 DBSCAN

DBSCAN is a clustering algorithm based on areas with high density of data points [73]. This clustering approach is achieved by initializing the algorithm with a threshold radius, commonly denoted ε and a threshold of the number of points which must be within the radius to form a cluster. With these parameters set, the algorithm proceeds to select data points and see if the desired number of data points can be found within the threshold radius. If the data point fulfills the criterion, the point is determined to be a core point from which a cluster is formed together with the neighboring points within the threshold radius. The algorithm iterates through the data points until all possible assignments have been made for data points. The data points which do not fulfill the criterion are considered and labeled as noise in the data.

4.5.3 Determining the optimal number of clusters

Due to the nature of K-means and other clustering approaches, Kodinariya and al. has reviewed approaches to determine the optimal number of clusters [74]. Among them, one of the most commonly used approach is the "elbow method". The elbow method is based on plotting cost function for different cluster numbers to identify the cost functions changes starts to flatten out. The plateau indicates that a further increase of the number of clusters will result in diminished improvements of the cost function. Often, the elbow method takes the distortion of the clusters as the cost function, which is computed by taking the sum of the euclidean distances from each point within clusters to the centers of their cluster [75]. Though it is commonly calculated by manually examining of the plot to find the elbow point, one library calculates the elbow point automatically [75]. The library utilizes the "kneedle" algorithm, which finds the point where the curvature of a graph starts to flatten after reaching a maximum curvature [76]. In this thesis, the Yellowbrick implementation of the kneedle algorithm will be used to determine the optimal number of clusters for K-means.

4.5.4 Implementation of clustering algorithms

With the preprocessed and embedded data as input, the clustering is performed using the K-means algorithm implemented in the Scikit-learn library [77]. The input is in the form of vectorized and scale-normalized representations of the preprocessed disengagement entries. Due to the nature of the data in the project, the exact number of clusters that is desired and results in the best clustering is not known to the user and some exploratory testing is necessary to identify the optimal number of clusters. Through every iteration, all the different vectorization methodologies are evaluated based on the cluster metrics that are achieved. In order to determine the optimal number of clusters, the elbow method is used as a way to determine when added clustering does not improve grouping of the data. The elbow method is originally a visual inspection of the plot of the mean distortion score to different number of clusters, where the elbow point is the position where the curve starts to flatten, indicating that an increase of the number of clusters leads to diminishing improvements. The implementation of the elbow method in Yellowbrick is used to determine the elbow point automatically for each iteration [78].

Chen *et al.* aims to find representatives of clusters by using a subset clustering approach [79]. They create an initial clustering of large groups and perform further clustering until they reach a subset of representatives of a desired amount. For the subset clustering, they opt for the usage of a medoid-based assignment of cluster centers instead of the centroid based assignment of centers. The difference is that medoid is based on assigning the center to the data point which has the highest similarity to the other data points as opposed to setting the cluster center based on finding the mean coordinates for the data points. Chen *et al.* reveals the benefit of using K-medoid in the possibility to get a cluster representative instantly in the form of the center point and also that the medoid approach leads to robustness against skewing of outliers in the dataset. The effectiveness of medoid selection for clustering categorical data is also reported by other studies searching for data mining approaches in vehicle datasets [3].

In this thesis, the cluster representatives are determined using a medoid-based approach. For each data point the pairwise distances are calculated in respect to all other data points and their sum is calculated. The data point which has the lowest total pairwise distance to the other data points is determined to be the cluster representative.

The implementation for DBSCAN is similar in many aspects and use the Scikit-learn library [77]. For DBSCAN clustering, the parameters needed include the minimum samples required to form a cluster and the ε-value which determines the radius in which samples needs to be placed to form clusters. The clustering is performed with varying values on these parameters to determine the optimal clustering.

4.6 Evaluation of clusters

In a study by Tomašev and Radovanović, tools are presented to determine the optimal parameters for clustering without analyzing the resulting clustering manually [80]. Also, the study describes a series of common metrics to evaluate clustering such as Davies-Bouldin index (DBI), the silhouette index and the Calinski-Harabasz index (CH index).

Another approach to evaluate clustering was reported by Sapna and Mohanty [81]. They compare the clustering results to the results achieved by randomly grouping data points to give an indication of how much potential improvement the clustering results in.

4.6.1 Silhouette index

Dudek present silhouette score as a clustering evaluation metric [82]. Silhouette score is a measurement of how similar objects within clusters are to each other. The silhouette score is calculated by first taking a data point and calculating the mean distance from the point to other data points within the cluster. Then, the mean distance from the same data point to data points in the nearest neighboring cluster is computed. The silhouette score is the difference of these two distances divided by the larger one of them and is a value between 1 and -1, calculated according to equation 4.1. The equation shows the formula to calculate the silhouette index. i is the index for a data point, N is the number of data points, $a(i)$ is the mean distance from point i to the other points in the same cluster, $b(i)$ is the smallest mean distance to all points in a different cluster. A value of 1 indicates that the data point is well matched to its own cluster and dissimilar to other clusters data points. A value of -1 indicates that the point has strong correlation to other clusters and weak cohesion with the cluster it is placed within.

$$\text{Silhouette} = \frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{4.1}$$

4.6.2 Calinski-Harabasz Index

Baarsch and Celebi presents the principle behind the Calinski-Harabasz (CH) index [83]. The CH index is used to measure the variance that exists between different clusters in comparison to the variance found within a cluster. First, the euclidean distance from the center of each cluster to the center of the complete dataset. Second, the mean euclidean distance from the center of the cluster to the data points within the cluster is calculated for each cluster. Thirdly, compute the within-cluster variance WC for each cluster as well as the between-cluster variance BC between the clusters. With these values calculated, the CH index is calculated according to equation 4.2. The equation shows the formula to calculate the CH index, where k

is the number of clusters and n is the numbers of data points clustered. A high CH index indicates strongly bound clusters with a high level of separation from other clusters.

$$CH = \frac{BC}{(k-1)} \cdot \frac{n-k}{WC} \tag{4.2}$$

4.6.3 Davies–Bouldin index

The Davies-Bouldin index (DBI) is a metric used to validate clustering performance [80]. Singh *et al.* uses the clustering to evaluate K-means clustering and finding the optimal numbers of clusters to sort similar groups of cereals [84]. They present DBI as a metric that aims to measure the compactness of clusters by calculating the distances between data points within clusters and comparing to the level of separation that the clusters have by calculating the distances between clusters, similar to the CH index. The difference lies in how these parameters are calculated. The algorithm takes the centroid of each cluster ($C(i)$) and the mean distance from all data points within a cluster to the centroid ($a(i)$) and maximize the value of $R(i)$ as defined by equation 4.3 by evaluating combinations of i and j where $i \neq j$ and i and j indicates the cluster number. The DBI is calculated according to 4.4 [80]. In other words, DB index takes pairwise difference while CH index takes the total variance within clusters in relation to the total variance between different clusters.

$$R(i) = \frac{a(i) + a(j)}{C(i) - C(j)} \tag{4.3}$$

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} R(i) \tag{4.4}$$

4.6.4 Implementation of evaluation

In this thesis, the evaluation of the clustering for K-means is based on evaluating mean silhouette scores, CH index, DBI as well as the optimal number of clusters using the elbow method based on the distortion score. This is used together with an estimation of how large portion of the disengagements the representative scenarios are able to accurately replicate as described in section 4.8.2. The clustering from DBSCAN can not be evaluated in the same way as K-means, but is also evaluated based on coverage of scenarios presented in section 4.8.2.

4.7 Visualizing the clustering result

In order to visualize the data of high dimensionality, the Yellowbrick clustering library is used together with Matplotlib in this thesis [78], [85]. The data points are both visualized using their inter cluster similarities, showing the similarities with

other clusters and as scatter plots. We use PCA to reduce the dimensions of the data. The cluster representations are visualized together with the cluster centers. The clustering results are also presented in CSV files that show which data points are to be considered border points.

4.8 Test scenario suite representation and coverage analysis

The metrics presented in section 4.6 serves as evaluation to indicate if well-formed clusters have been found. However, the main goal is to see if the clustering is a viable way to efficiently test the dataset. In order to evaluate if the resulting clustering representatives are good representations of the dataset, the coverage of scenarios based on the representations needs to be evaluated.

Due to this problem of determining if the clustering has achieved the goal to group similar entries in the same clusters in the context of disengagement scenarios and not only their vector representations, a manual evaluation is also performed. The evaluation is based on the process of parameterizing and capturing the scenarios in a standardized manor.

4.8.1 Test scenario representation

In the California DMV dataset, concrete values are not included and due to the formulation of disengagement scenarios in the California DMV dataset, concrete values are excluded and functional scenarios provide too little detail to distinguish different scenarios from each other. Therefore, the logical scenarios will be used to parameterize the disengagement. An example is the disengagement entry "Camera vision impeded by sun. Vehicle not in an active construction zone. No emergency vehicles or collisions present in the vicinity. Weather and/or road conditions dry in the area.". In order to recreate this scenario, the relevant characteristics are placed in the columns, and their values are assigned in the rows as shown in table 4.3.

Sunshine on camera	In construction zone	Emergency vehicles or collisions present	Weather and road conditions
True	False	False	Dry

Table 4.3 The table shows the parameterization of the disengagement scenario "Camera vision impeded by sun. Vehicle not in an active construction zone. No emergency vehicles or collisions present in the vicinity. Weather and/or road conditions dry in the area.".

This way of parameterization makes it possible to describe the disengagements in a way that makes it possible to compare disengagement scenarios by comparing the parameter values shared between disengagements. Further, it also makes it possible to make combinations of parameters to generate test scenarios not present in the dataset by combining parameter values which can co-exist between different

scenarios. The result is a test scenario suite containing already existing disengagement scenarios and scenarios not found in the disengagement data base. The extraction of the disengagement conditions is determined manually. Parameterization is made for each cluster based on the data point that is the representative of that cluster.

4.8.2 Coverage evaluation

The coverage of the dataset is calculated by using the parameterization described in 4.8.1 to the filtered dataset. The parameterized versions of the entries are used to evaluate the level of coverage that have been achieved by clustering and selecting clustering representatives.

First, the number of parameter values found in the cluster representatives is determined. The total parameter value coverage is then calculated by calculating the proportion of the total number of parameter values in the dataset have been captured by the representatives.

Secondly, the number of clusters have a great effect on the number of parameter values that can be covered. This is due to the number of cluster representatives increasing and therefore results in a higher possible number of parameter values covered by them. Therefore the parameter values found with the cluster representatives will also be compared to the number of parameter values that would be captured with optimal scenario selection. The optimal coverage is calculated by combining the scenarios that achieve the highest number of unique parameter values when combined.

Another coverage value is achieved by selecting cluster representatives at random, which shows how much the selection process has improved the representative selection over random assignment.

For each parameterization, the basis will be to let parameters which can't be combined and have similar scope to be grouped together. One such example is the parameter group 'autonomous vehicle location' which can have values such as 'Intersection' or 'Construction site', which have values that are related in the sense that they give the location of the vehicle. But they are not possible to combine as the vehicle must be at one place in a certain time. The parameterization is performed manually by the author on the non clustered dataset. The parameters are never passed to the clustering model and are not used when determining the cluster coherence's of data points.

5

Evaluation

To determine the effectiveness of the scenario selection methodology, we will investigate its cluster quality and coverage performance. In section 5.1, we present the parameters found using the parameterization technique described in section 4.8.1. Moving to section 5.2 we evaluate to what extent the cluster representatives selected with K-means can cover the total number of parameters in the dataset set. The cluster quality is also evaluated using silhouette scoring, CHI and DBI. In 5.3 we also evaluate the coverage of parameters as well, but for the DBSCAN based selection methodology.

We investigate RQ1 by determining the level that the embeddings can separate disengagements both visually and using the coverage results. A high level of spread in the representations combined with a high coverage result indicates that the embeddings can represent similarities for the disengagements. We also investigate RQ2 by determining if we can achieve well-defined clusters indicating that the data can be clustered and achieve good representative selection. This is done by investigating the performance in terms of silhouette score, CHI and DBI where well-scoring measurements indicate effective clustering. We can link the findings in terms of parameters with RQ3, as they show if the parameterization methodology is able to describe the scenarios in a standardized way. Lastly, RQ4 concerning the coverage performance of the methodology is investigated with the coverage scores presented in sections 5.2 and 5.3. This reveals what the highest level of coverage performance the methodology can achieve is.

5.1 Parameterization of scenarios

I present the results from the parameterization process described in 4.8.1 in this section. The parameters necessary to describe all the disengagements in the dataset fully is presented in tables 5.1 and 5.2. The parameter values that are mutually exclusive (right column of the tables) in the dataset are grouped within a certain parameter group (left column of the tables). This classification is devised by me, as a way to better show how the parameters can be combined to craft disengage-

ment scenarios. The total number of parameter groups is 27 and the total number of parameter values is 92.

Autonomous Vehicle Scenario Parameters (Part 1)	
Autonomous Vehicle Road Blockage	Construction cones, Construction work, Throughlanes fully blocked, Garbage on road, Road debris
Autonomous Vehicle Vision	Limited view of oncoming lane, Camera impeded
Sight Blocker	Dust, Sun, Actor
AV Location	Intersection, Turn, Roundabout, Open railway, Junction, Street, Parking lot
Back Wall Modifier	Close to back wall
Side Wall Modifier	Close to side wall
Curb Modifier	Curb nearby
Lane Position Modifier	Over lane boundary, Off center in lane, At lane marker
Road Curve Modifier	Straight road
Intersection Modifier	3-way, 4-way
Crosswalk Modifier	Crosswalk
Width Modifier	Narrow path, Wide lane
Bike Lane Modifier	Bikelane present
Sign Modifier	Stop sign, Traffic light, Pedestrian yield sign, Construction warning, Stop line
Traffic Light Modifier	Green, Red Yellow, Green turn
Autonomous Vehicle Rights	Right of way
Autonomous Vehicle Limitation	Double yellow line
Autonomous Vehicle Trajectory	Forward, Backward, Turn, Turn over oncoming traffic lane, Merge, Start after full stop, Brake
Merge Destination	Highway, Lane, Express-lane

Table 5.1 The table shows the parameters used to construct the scenarios present in the dataset. The parameter groups is shown in the left column and their values are shown in the right column. (Part 1).

Autonomous Vehicle Scenario Parameters (Part 2)	
Actor Type	Vehicle, Big vehicle, Motorcycle, Cyclist, Pedestrian, Traffic, Bus, Animal, Construction vehicle, No traffic, Pedestrians
Actor State	Parked, Oncoming traffic, Double parked, Sitting in parked car, Private driveway, Fire lane, Outer lane, Parked on both sides, Parking lane, In target lane, Side of road
Actor Position towards Autonomous Vehicle	In front of, Behind of, Adjacent lane
Actor Action	Merge to AV lane, Open door in AV trajectory, Start from stop, Cross street, Leave car, U-turn, Turn, Turn passing AV trajectory, Enter oncoming traffic lane, Go forward, Yield to other actor, Braking, Reversing, Cut-in, Swerve, Collide with AV, Stand still
Collision Present	Contact with AV
Actor Intent Legal/Odd/Illegal	Illegal, Legal, Unexpected
Weather	Sunny, Cloudy
Road Conditions	Dry

Table 5.2 The table shows the parameters used to construct the scenarios present in the dataset. The parameter groups is shown in the left column and their values are shown in the right column. (Part 2).

5.2 Coverage achieved with K-means clustering

As shown in tables 5.3 through 5.4. The results from clustering without dimension reduction show that no approach achieve significantly higher coverage than random representative selection would. With the lowest achieving approach performing at approximately 89 % of a random assignment and the highest at 102 %. Considering that the silhouette score, DBI and CH index achieved for these measurements is relatively low, the clustering is not able to achieve well-formed clusters with the embeddings. This could be one of the reasons for the low coverage improvements over random selection.

Table 5.3 The table shows the silhouette score, DBI and CH index achieved with the different embedding methods. The optimal number of clusters as determined by the elbow method is also presented.

Embedding	K	Sil.	DBI	CHI
word2vec	12	0.1	2.3	9.95
word2vec (Google news)	14	0.01	1.71	6.06
doc2vec	6	0.12	1.67	67.13
spaCy	12	0.1	2.18	9.42
TF-IDF	13	0.09	2.86	5.63

Table 5.4 The table shows the coverage results achieved for K-means clustering with the embeddings. The first two columns show the embedding used and the optimal number of clusters determined by the elbow method. The next four columns display the coverage of parameter, parameter values, the coverage relative to optimal selection and the coverage relative to random assignment.

Embedding	K	Param. cov.	Val. cov.	Opt. cov.	Rand. cov.
word2vec	12	55.17%	26.09%	60%	101.61%
word2vec (Google news)	14	51.72%	28.26%	59.09%	99.24%
doc2vec	6	34.48%	15.22%	60.87%	98.61%
spaCy	12	55.17%	22.83%	52.5%	88.87%
TF-IDF	13	41.38%	27.17%	59.52%	100.22%

The measurements in tables 5.3 and 5.4 is also visualized. We present the data with 2-dimensional scatter plots in figures 5.1 through 5.5. We assigned colors to the data points to show cluster membership. To achieve a 2D representation of the data, we used PCA to reduce the dimension of the embeddings.

Figure 5.1 shows the disengagements embedded with doc2vec assigned to 6 clusters with K-means. Most of the data points are within the horizontal range of 0 to 0.4. This indicates that most of the data are represented similarly with the embeddings. The plot reveals that fewer data points have been assigned to the two clusters to the far right. In the two clusters, the data points also appear more spread-out. This indicates that the two clusters to the far right can be groupings of outliers, considering their internal variance and the separation from the other clusters. Each cluster is clearly separated in the horizontal axis. This indicates that clustering primarily is based on a small selection of features. The grouping of the data points as a whole suggests that the doc2vec embeddings may not be able to separate the disengagements in the dataset.

Figure 5.2 shows the representation of the disengagements embedded with spaCy assigned to 12 clusters. The clusters appear well separated in the outer edges of the data with some overlapping in the middle of the plot. The lack of separation can reveal a problem in terms of clustering. This indicates that the disengagements

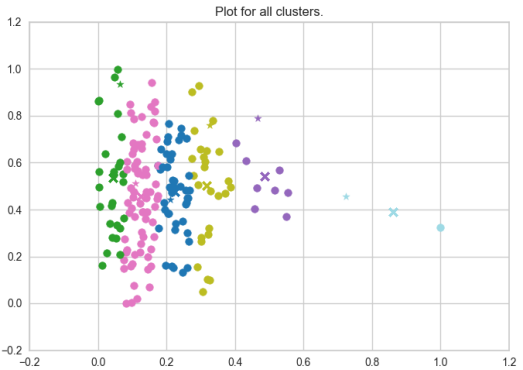

Figure 5.1 The figure shows the plot for the K-means approach using doc2vec embedding performed without dimensionality reduction. The stars show the cluster representatives, the crosses show centroids of clusters.

are hard to clearly divide into clusters. The results in table 5.3 show a high DBI, a low silhouette score and a low CH index for the measurement as well, indicating low levels of separation. The data does however appear well separated, which indicates that the embeddings are able to show differences in disengagements. Large scattering could also be the reason for the low clustering results, as data points are seen as more dissimilar. The spread in both axes in the graph of clusters also indicate that more features carry information relevant for clustering. The disengagements are varied in their natural language representations and therefore a utilization of a high number of features to separate disengagements is expected.

Figure 5.3 shows the representation of the disengagements embedded with TF-IDF assigned to 13 clusters. The two clusters to the far right and at the top are well defined and separated, while the other clusters are less separated. This lack of separation is also indicated in figure 5.3, with a low silhouette score, low CH index and a high DBI for the measurement.

Figure 5.4 shows the representation of the disengagements embedded with word2vec (Google news) assigned to 14 clusters. The data is tightly bound together with the exception for 5 data points that are more spread out and form their own clusters. The low level of separation between most data points indicate that the majority of the disengagements are represented similarly with the embeddings. The word2vec (Google news) model is built from a broad data base from many fields, which could be the reason that the overarching topic of autonomous vehicles cause the disengagements to be represented as highly similar.

Figure 5.5 shows the representation of the disengagements embedded with word2vec assigned to 12 clusters. The data shows a large spread. Some fairly well-separated clusters can be seen, but also a high level of overlap in the middle of the

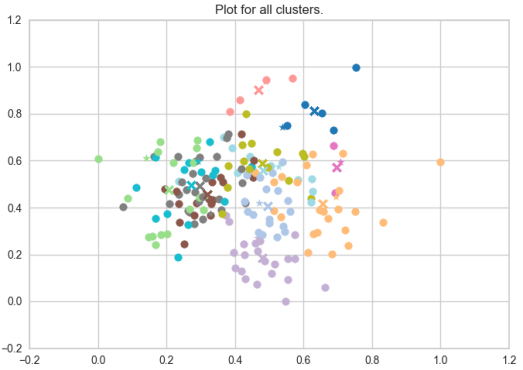

Figure 5.2 The figure shows the plot for the K-means approach using spaCy embeddings performed without dimensionality reduction. The stars show the cluster representatives, the crosses show centroids of clusters.

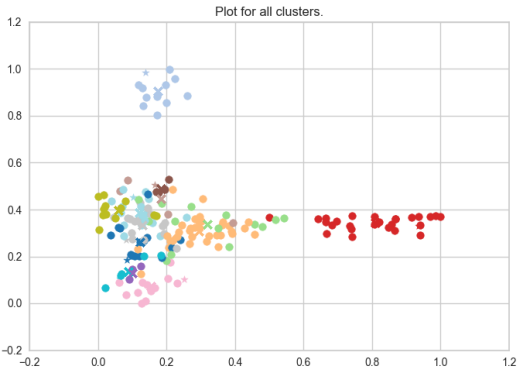

Figure 5.3 The figure shows the plot for the K-means approach using TF-IDF embeddings performed without dimensionality reduction. The stars show the cluster representatives, the crosses show centroids of clusters.

plot and its upper left corner. The spread indicates that the embeddings can represent differences in disengagement scenarios which makes the clustering approach feasible. However, the overlap between clusters may effect the accuracy. Many data points may be placed on the border between clusters and have weaker associations with a specific cluster.

The summary of the results from figures 5.1 through 5.5 will be made. The clustering with the embeddings in their original format gives some insight to their effec-

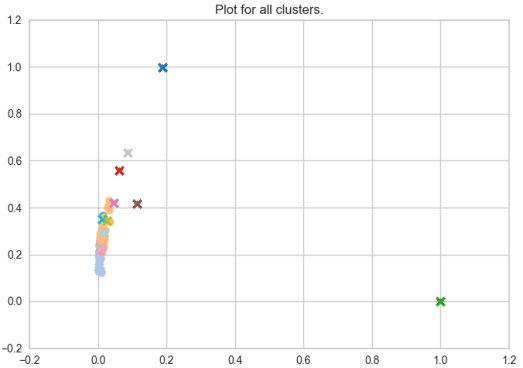

Figure 5.4 The figure shows the plot for the K-means approach using word2vec (Google news) embeddings performed without dimensionality reduction. The stars show the cluster representatives, the crosses show centroids of clusters.

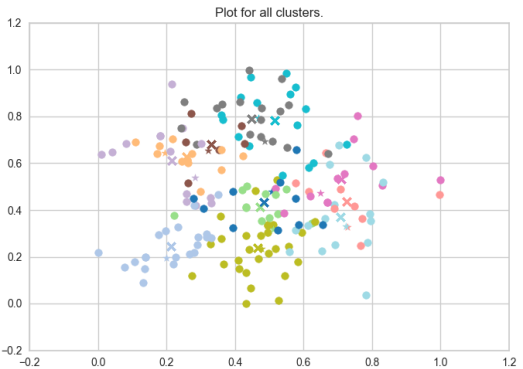

Figure 5.5 The figure shows the plot for the K-means approach using word2vec embeddings performed without dimensionality reduction. The stars show the cluster representatives, the crosses show centroids of clusters.

tiveness. The visual inspection indicates that the spaCy and word2vec embeddings can represent the disengagements well separated. For word2vec (Google news), TF-IDF and doc2vec the data appears with quite a low level of separation, indicating that they may not be able to capture the differences between disengagement scenarios. However, the metrics presented in table 5.3 does warrant further evaluation as all embedding approaches achieve low silhouette scores, high DBI and low CH index. Which indicates that clustering using the embeddings will not form well-

defined clusters.

Now we will present the results achieved with PCA dimensionality reduction applied to the embeddings before clustering. The embeddings were reduced to dimensions between 2 and 99. The values were initially tested with increments of 10ths in dimensions. When the range that high scoring results were present in became apparent, increments of one were used to find the optimal dimensionality. We determined the optimal number of clusters (K) with the elbow method based on distortion scoring. The tested number of clusters were in the range of 2 to 19. We present the highest scoring measurements in tables 5.5 through 5.11

Tables 5.5 through 5.7 concern the clustering metrics in the form of silhouette scores, CH indexes and DBI. These tables gives us some insight into what manipulation of the input data leads to high-scoring clusters. For all measurements except for the word2vec (Google news) approach performance in DBI, the highest results were achieved for embeddings in two dimensions. The high scores indicate that similar data points have been grouped together and dissimilar data points are well-separated. This indicates that the data points can be separated most distinctively when the clustering is based only on the most predominant features. 4 embedding approaches scored the highest in all three metrics with the same number of dimensions and clusters. word2vec (Google news) did however achieve its highest DBI measurement with a higher dimensionality together with a higher number of clusters. In regards to the performance seen in the tables 5.5 through 5.7, the word2vec (Google news) and TF-IDF embeddings outperform word2vec, doc2vec and spaCy in terms of silhouette score, CH index and DBI. This indicates that in regards to these clustering metrics, the TF-IDF and word2vec (Google news) are the best embedding approaches for achieving well-defined clusters. The word2vec (Google news) achieved the best scoring and the TF-IDF approach achieving the second best results in all three metrics.

Table 5.5 The table shows the measurements that achieved the highest silhouette score for each embedding methodology. It is presented together with the number of dimensions that the vectorized data was reduced to as well with the number of clusters that the score was achieved for.

Embedding	N dim.	K	Silhouette score
word2vec	2	8	0.39
word2vec (Google news)	2	6	0.56
doc2vec	2	7	0.37
spaCy	2	6	0.38
TF-IDF	2	5	0.52

Table 5.6 The table shows the measurements that achieved the highest Calinski-Harabasz index for each embedding methodology. It is presented together with the number of dimensions that the vectorized data was reduced to as well with the number of clusters that the score was achieved for.

Embedding	N dim.	K	CH Index
word2vec	2	8	162.98
word2vec (Google news)	2	6	591.25
doc2vec	2	7	165.32
spaCy	2	6	136.66
TF-IDF	2	5	434.5

Table 5.7 The table shows the measurements that achieved the lowest Davies-Bouldin index for each embedding methodology. It is presented together with the number of dimensions that the vectorized data was reduced to as well with the number of clusters that the score was achieved for.

Embedding	N dim.	K	DBI
word2vec	2	8	0.81
word2vec (Google news)	8	9	0.32
doc2vec	2	7	0.8
spaCy	2	6	0.81
TF-IDF	2	5	0.57

In table 5.8, the measurements that achieved the highest level of coverage based on the parameters defined in section 4.8.1 is presented. The table shows that all of the approaches achieve the highest result for the maximum number of possible clusters, 19 except for the TF-IDF measurement. This means that a large number of cluster representatives selected leads to a high coverage. This is to be expected as the number of possible parameters that can be covered is increased. The dimensionality of the data was also high, ranging between 20 and 43 dimensions for the measurements. A reason for this can be that a higher number of dimensions makes it easier to interpret disengagements numerically as dissimilar. The optimal coverage measurements were achieved with a higher number of clusters and dimensions than those that scored the highest in silhouette score, CH index and DBI. This could indicate that a higher level of features can achieve a higher level of variation between the clusters which in turns increases the number of parameters that were covered. However, the scores achieved were not very high, with a range between 65%-69% coverage for the parameters with these measurements. word2vec (Google news), doc2vec and spaCy outperforms word2vec and TF-IDF in these measurements by a small margin of approximately 4 percentage points.

Table 5.8 The table shows the measurements that achieved the highest coverage of parameters for each embedding methodology. It is presented together with the number of dimensions that the vectorized data was reduced to as well with the number of clusters that the score was achieved for.

Embedding	N dim.	K	Coverage of parameters
word2vec	43	19	65.52%
word2vec (Google news)	25	19	68.97%
doc2vec	41	19	68.97%
spaCy	37	19	68.97%
TF-IDF	20	16	65.52%

Similar results can also be seen in table 5.9, which shows the highest coverage of parameter values achieved. The value coverage is within the ranges 38%-41% for the measurements, which is quite low. Like for the parameters, a high number of clusters with a high dimensionality of input data achieved a high value coverage. This can be explained with the same reasoning regarding the benefits of a high number of representatives and an improved ability to separate dissimilar disengagements. One difference between tables 5.8 and 5.9 is that the TF-IDF achieves a higher coverage of values with 19 clusters. Where the highest parameter coverage is achieved with 16 clusters. The dimensionality of the input data between the two tables also have slight differences, with the spaCy and TF-IDF measurements changing from 37 to 58 and 20 to 36 respectively from table 5.8 to 5.9.

Table 5.9 The table shows the measurements that achieved the highest coverage of the values within parameters for each embedding methodology. It is presented together with the number of dimensions that the vectorized data was reduced to as well with the number of clusters that the score was achieved for.

Embedding	N dim.	K	Coverage of values
word2vec	42	19	41.3%
word2vec (Google news)	25	19	39.13%
doc2vec	41	19	39.13%
spaCy	58	19	39.13%
TF-IDF	36	19	38.04%

We also evaluate the value measurements in regards to what could be achieved with an optimal selection of cluster representatives. These measurements are presented in table 5.10. The measurements have a high spread in terms of dimensionality and the number of clusters. The highest results were achieved for doc2vec, scoring approximately 85%, with the other measurements within the range of 71% to 74%. The high scoring of the doc2vec approaches indicate that the representative

scenarios selected are well-spread and have parameter values that to a higher degree differ from each other. Considering that the number of representatives selected were fairly low (7), this indicates that the initial representative choices are made with higher accuracy. This means that the doc2vec approach is the best in terms of a high level of coverage per cluster representative selected.

Table 5.10 The table shows the measurements that achieved the highest coverage of the values within parameters in relation to what coverage is achievable based on combining an optimal selection of scenarios for parameter value coverage. The results are presented for each embedding methodology. It is presented together with the number of dimensions that the vectorized data was reduced to as well with the number of clusters that the score was achieved for.

Embedding	N dim.	K	Compared to optimal selection
word2vec	42	19	71.7%
word2vec (Google news)	17	11	73.68%
doc2vec	6	7	84.62%
spaCy	2	6	69.56%
TF-IDF	4	6	73.91%

Lastly, measurements were made to see what improvement could be made over random selection of cluster representatives. This result is presented in table 5.11. The highest scoring measurements share many similarities with table 5.10. One difference is that the dimensionality and cluster numbers changed for spaCy from 2 to 6 and 6 to 8 respectively. The highest results were instead by doc2vec and word2vec (Google news) with approximately 138% respectively 126% achieved coverage compared to random selection. These results reveal that the embeddings can outperform random selection, with the doc2vec approach showing the highest improvement.

Table 5.11 The table shows the measurements that achieved the highest coverage of the values in relation to what coverage is achievable based on combining scenarios at random. The results are presented for each embedding methodology. It is presented together with the number of dimensions that the vectorized data was reduced to as well with the number of clusters that the score was achieved for.

Embedding	N dim.	K	Compared to random selection
word2vec	42	19	119.5%
word2vec (Google news)	17	11	125.94%
doc2vec	6	7	137.72%
spaCy	6	8	113.09%
TF-IDF	4	6	119.72%

The results in tables 5.8 through 5.11 is summarized in 5.12. In total, the result shows us that the word2vec (Google news) approach led to the highest scoring results in terms of silhouette score, DBI and CH index. This result indicates that the approach is the optimal selection to achieve accurate clustering. However, the coverage evaluations reveals that the word2vec (Google news) approach only was able to perform the best in terms of parameter coverage in shared first place. When it came to coverage of the values, the word2vec approach performed slightly better than the other approaches. In regards to coverage in relation to the maximum achievable for the approach, doc2vec achieved the highest results. Lastly, in terms of improvement over random selection, the doc2vec approach had the highest performance. This results leads to inconclusive results regarding which embedding approach is best for representative selection for disengagement scenarios. However, it can be seen that all approaches can be tuned in a way that outperforms random selection. But considering the current maximum levels of coverage that have been achieved, the approach does not select representatives in a way that is close to achieving 100% coverage of all parameters and all values in the dataset.

Table 5.12 The table shows a compilation of the highest scoring measurements in regards to silhouette score, DBI, CH index and the coverage evaluation values described in section 4.8.2

Metric	Vec.	N dim.	N clusters	Value
Sil. score	word2vec (Google news)	2	6	0.56
DBI	word2vec (Google news)	8	9	0.32
CH Index	word2vec (Google news)	2	6	591.25
Param. cov. (1)	doc2vec	41	19	68.97%
Param. cov. (2)	word2vec (Google news)	25	19	68.97%
Param. cov. (3)	spaCy	37	19	68.97%
Value cov.	word2vec	42	19	41.3%
Compared to optimal selection	doc2vec	6	7	84.62%
Compared to random selection	doc2vec	6	7	137.72%

The measurements shown in table 5.12 is also visualized. For the entries which

scored highest in regards to the clustering evaluation metrics, which can be seen in figures 5.6 through 5.12.

Figure 5.6 shows the cluster measurement that achieved both the highest silhouette score and CH index. The plot shows that the clusters are well defined, with the outlying data points being assigned their own clusters. As discussed in regards to figure 5.4, a lot of the data points are placed in a narrow interval meaning that different disengagements may be hard to differentiate with this embedding.

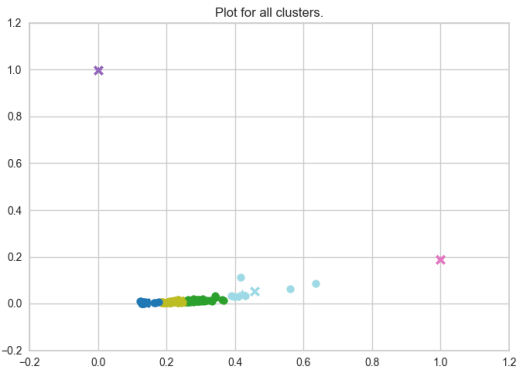

Figure 5.6 The figure shows a scatter plot of the K-means clustering with the highest silhouette score and CH index. The embedding was word2vec (Google news) reduced to 2 dimensions, with 6 cluster formed, represented by different colors. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

Figure 5.7 shows the clustering that achieved the highest DBI. The clusters are well-separated and outliers are assigned to own clusters. The spread of the data points indicate that differences between data points can be discerned.

Figure 5.8 shows the clustering which achieved the highest level of value coverage. The clusters appear to be overlapping with a low level of separation. The high level of value covered seem to be an effect of the high separation between data points and the forming of many clusters as opposed to forming clearly defined clusters.

Figures 5.9 through 5.11 show the two measurements that achieved the highest level of parameter coverage. Both embeddings make the points well separated. However, similarly to the embedding that scored the highest in terms of value coverage shown in figure 5.8, many clusters have been formed but show low levels of separation.

Figure 5.12 shows the measurement that achieved the highest improvement compared to random disengagement representative selection. The two clusters to the far right show a fairly high level of separation with the rest of the data points,

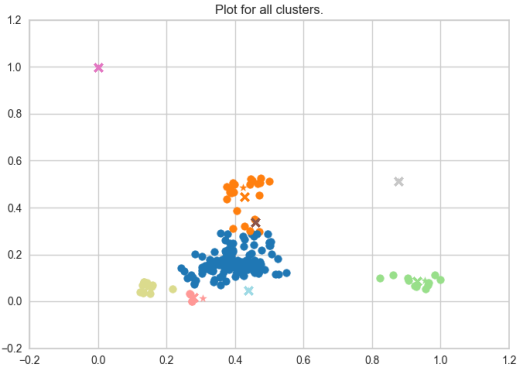

Figure 5.7 The figure shows a scatter plot of the K-means clustering with the lowest DBI. The embedding was word2vec (Google news) reduced to 8 dimensions, with 9 cluster formed, represented by different colors. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

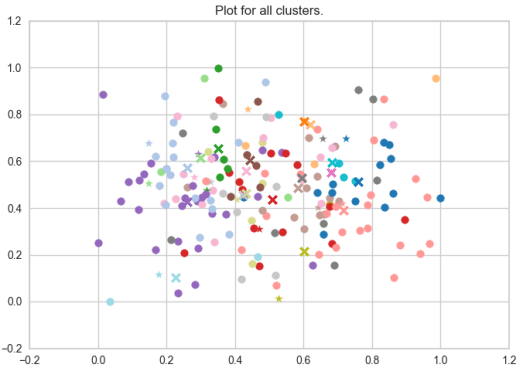

Figure 5.8 The figure shows a scatter plot of the K-means clustering with the highest coverage of parameter values. The embedding was word2vec reduced to 42 dimensions, with 19 cluster formed, represented by different colors. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

with some overlap in the center of the plot. The spread of the data points indicate that differences can be determined with the embeddings. The overlapping of the clusters indicate that the assignments made may be uncertain as the data points are close to being assigned to another cluster.

The combined results of the scatter plots as well as tables indicate that no clear correlation between well-formed cluster and high coverage can be established. Ta-

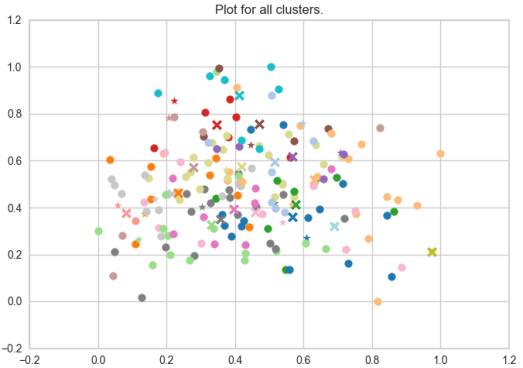

Figure 5.9 The figure shows a scatter plot of the K-means clustering with the highest coverage of parameters. The embedding was spaCy reduced to 37 dimensions, with 19 cluster formed, represented by different colors. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

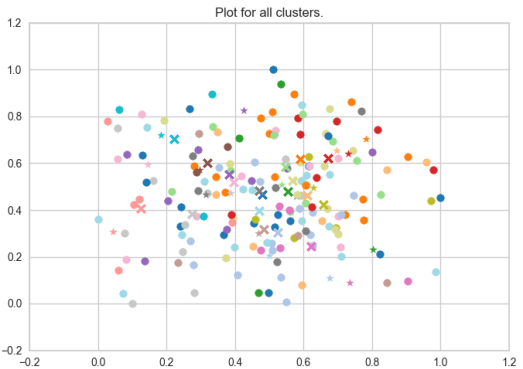

Figure 5.10 The figure shows a scatter plot of the K-means clustering with the highest coverage of parameters. The embedding was doc2vec reduced to 41 dimensions, with 19 cluster formed, represented by different colors. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

ble 5.12 and figures 5.9 through 5.12 show that the highest scoring measurements in terms of coverage of parameters and values are not associated with clear cluster formation in the figures. The best performing clustering in terms of cluster clarity was shown for the measurements associated with a high CH index, low DBI and high silhouette score as shown in figure 5.6 and figure 5.7. These figures did however show that the embeddings may not be able to separate different disengagements

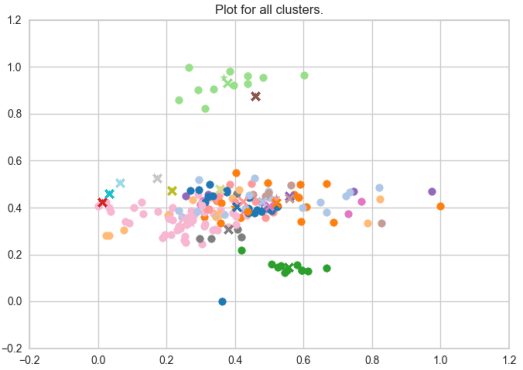

Figure 5.11 The figure shows a scatter plot of the K-means clustering with the highest coverage of parameters. The embedding was word2vec (Google news) reduced to 25 dimensions, with 19 cluster formed, represented by different colors. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

as most data points are placed with low separation. Though these results together indicate that the selection of disengagement representatives using K-means is unsuccessful, the comparison with random assignments show some promising results. The measurement that achieved the highest performance over random assignment did show some level of cluster separation in figure 5.12. This may indicate that most of the embeddings can be used to achieve results higher than random selection.

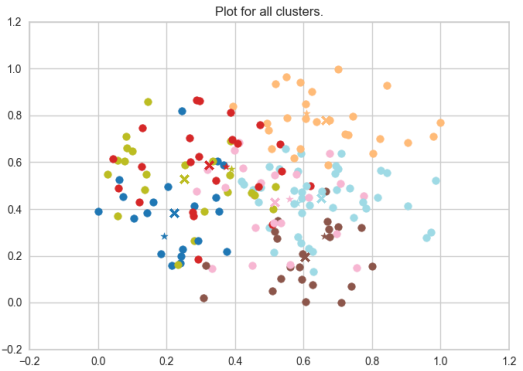

Figure 5.12 The figure shows a scatter plot of the K-means clustering with the highest coverage of values in relation to random selection and optimal selection. The embedding was doc2vec reduced to 6 dimensions, with 7 cluster formed, represented by different colors. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

5.3 Coverage achieved with DBSCAN clustering

This section deals with the results from the DBSCAN clustering. The difference in clustering between K-means and DBSCAN means that the silhouette score, DBI and CH index can not be used to evaluate the clustering. However, the coverage metrics will be presented the same way as they were for K-means. The measurements were achieved by clustering with a minimum samples parameter set within a range of 2 to 6 samples. The ε value was empirically tested to achieve the best scoring measurements. DBSCAN does not instantiate with a set number of clusters. Therefore, the optimal number of clusters will be determined by the DBSCAN algorithm and not by the knee method. The resulting coverage of the dataset is presented in tables 5.13 through 5.16.

Table 5.13 presents the highest parameter coverage achieved. The measurements have a low level of dimensions, ranging from 3-5 with word2vec (Google news) being an exception which scored its highest measurement for 20 dimensions. The number of clusters formed were within the range 14-35 which, in comparison to the limit set on K-means of 19, is high. The minimum samples for each measurement were 2. This shows that the measurements scored high when a low number of data points were needed to see a region as dense and form clusters. This in turn leads to the ability to form more clusters and therefore reach a higher coverage. The coverage results have a high level of spread between 55% for doc2vec embeddings and 80% with spaCy embeddings. Compared to the parameter coverage in the K-means approach, the results indicate that the DBSCAN approach achieves higher coverage. However, the DBSCAN measurements did not have an upper limit and the increased coverage could be attributed to the difference between the highest DBSCAN measurement with 35 clusters and the K-means measurement with 19 clusters.

Table 5.13 The table shows the measurements that achieved the highest coverage of parameters for each embedding. The measurements are presented with the number of dimensions that the input was reduced to, the number of clusters that were formed and the minimum samples parameter value for DBSCAN.

Embedding	N dim.	N clusters	Coverage	Min samples
word2vec	5	19	72.41%	2
word2vec (Google news)	20	18	62.07%	2
doc2vec	3	14	55.17%	2
spaCy	4	35	79.31%	2
TF-IDF	3	31	68.97%	2

Moving on, the highest measurements for coverage of values is presented in 5.14. The value coverage is within the ranges of 26%-53% for the DBSCAN approach. The dimensionality of the data that resulted in the highest measurements are similar to the ones that achieved the highest parameter coverage. The only

differences being the word2vec and word2vec (Google news) measurements. The word2vec measurement changed from 5 to 3 dimensions and the word2vec (Google news) measurement changed from 20 to 25 dimensions. They also were the only one with a different number of clusters, changing from 19 to 26 for word2vec and from 18 to 23 for word2vec (Google news). The performance of the approaches still ranked in the same ordering as the parameter coverage, with the spaCy remaining the embedding with the highest coverage and doc2vec remaining the lowest measurement.

Table 5.14 The table shows the measurements that achieved the highest coverage of values for each embedding. The measurements are presented with the number of dimensions that the input was reduced to, the number of clusters that were formed and the minimum samples parameter value for DBSCAN.

Embedding	N dim.	N clusters	Cov.	Min samples
word2vec	3	26	46.74%	2
word2vec (Google news)	25	23	40.22%	2
doc2vec	3	14	26.09%	2
spaCy	4	35	53.26%	2
TF-IDF	3	31	44.65%	2

The measurements for the embedding approaches in order to evaluate how they performed in relation to the optimal selection of cluster representatives is presented in table 5.15. The measurements achieve fairly high coverage in this regard except for the doc2vec approach. doc2vec achieves a maximum of 56% of the optimal selection coverage, while the other achieves scores of 75% and above. The highest scoring is word2vec (Google news) and spaCy, both with a coverage of 79%. A notable difference in the higher scoring measurements in this table in comparison to table 5.13 and 5.14 is that the results are achieved for lower number of clusters. The dimensionality of the data is also lower in the measurements for all approaches except for the doc2vec and TF-IDF measurements where the dimensionality remained 3 throughout the measurements. This could be attributed to the fact that the initial assignments have low chance of picking representatives with overlapping parameters and therefore scoring closer to the optimal selections. One difference is also that the minimum samples to form clusters where higher for the doc2vec and spaCy measurements that scored highest in table 5.15. The change of the minimum sample parameter indicates that selections made from regions with high density can represent a larger number of disengagements.

Table 5.15 The table shows the measurements that achieved the highest coverage of parameter values in relation to what could be achieved with optimal selection. The measurements are presented with the number of dimensions that the input was reduced to, the number of clusters that were formed and the minimum samples parameter value for DBSCAN.

Embedding	N dim.	N clusters	Compared to optimal selection	Min samples
word2vec	2	2	75%	2
word2vec (Google news)	12	5	78.95%	2
doc2vec	3	6	56.52%	3
spaCy	3	5	78.95%	3
TF-IDF	3	6	76.92%	2

Lastly, the evaluation of how the approaches performed compared to random selection is presented in table 5.16. The table reveals that a very low number of clusters are associated with a high improvement over the random cluster representation selection. With no approach forming more than 3 clusters. The dimensionality of the data was also low with all approaches performing best with embeddings of 3 dimensions or less. The word2vec and TF-IDF approaches increased the minimum number of samples necessary to form clusters while no embedding approach decreased the parameter. These combinations indicate that the highest level of improvement in comparison to random selection is achieved when clusters are formed from regions which are dense. As the number of clusters are quite low for the measurements achieved, this could indicate that there are a few dense regions where cluster representations accurately represent a large portion of the data.

Table 5.16 The table shows the measurements that achieved the highest coverage of parameter values in relation to what could be achieved with random selection. The measurements are presented with the number of dimensions that the input was reduced to, the number of clusters that were formed and the minimum samples parameter value for DBSCAN.

Embedding	N dim.	N clusters	Compared to random selection	Min samples
word2vec	2	1	165.07%	4
word2vec (Google news)	2	2	140.72%	2
doc2vec	2	2	105.43%	3
spaCy	3	3	136.12%	3
TF-IDF	3	3	160.97%	4

In order to evaluate the clustering results from DBSCAN as a whole, the best scoring results from each table is presented in table 5.17. In total, the result reveals that the DBSCAN approach is able to achieve a high improvement over random sce-

nario selection with the best scoring approach, word2vec, achieving 165% coverage compared to random selection. The value coverage is fairly low for the approach, with the spaCy approach reaching 53% coverage but capturing a relatively high number of parameters in the dataset with 79% parameter coverage. Finally, the coverage compared to what is realistically achievable is fairly high for all approaches except doc2vec as can be seen in table 5.15 with a coverage of 56% of the highest achievable. Both the word2vec (Google news) and spaCy approaches achieved high coverage in this regard, scoring 79% for 5 clusters formed.

Table 5.17 The table shows the measurements that scored highest in tables 5.13 through 5.16. The measurements are presented with the number of dimensions that the input was reduced to, the number of clusters that were formed and the minimum samples parameter value for DBSCAN.

Metric	Embedding	N dim.	N clusters	Min samples	Value
Coverage of parameters	spaCy	4	35	2	79.31%
Coverage of values	spaCy	4	35	2	53.26%
Compared to optimal selection (1)	word2vec (Google news)	12	5	2	78.95%
Compared to optimal selection (2)	spaCy	3	5	3	78.95%
Compared to random selection	word2vec	2	1	4	165.07%

The results presented in table 5.17 is visualized in figures 5.13 through 5.16. Figure 5.13 shows the measurement that achieved the highest coverage for parameters and parameter values. The points that have been labeled as noise are visualized in black. The figure has a large number of noise points while the clusters formed are fairly small and are well-separated. The number of clusters are high, with 35 cluster representatives selected being approximately 19% of the full size of 184 disengagements clustered. This result indicates that the clustering approach needs to select a high number of disengagement representatives to achieve a high coverage of the dataset.

Figure 5.14 and figure 5.15 show the measurements that achieved the highest coverage in relation to what optimal representative selection would result in. For figure 5.14, only a few outliers are labeled as noise points as well as some points in the center. The clustering has formed clearly separated clusters with some overlapping in the middle of the plot. In figure 5.15, the data has a greater spread and

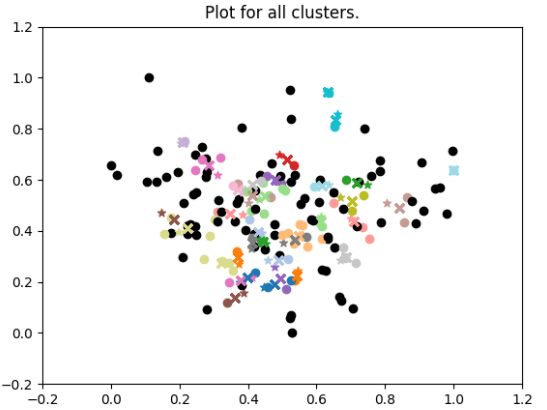

Figure 5.13 The figure shows a scatter plot of the DBSCAN clustering with the highest parameter coverage and parameter value coverage. The embedding was spaCy reduced to 4 dimensions, with 35 clusters formed, represented by different colors. The noise points are black. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

the clusters are more overlapping. The low level of clusters for both figures indicate that assignment of cluster representatives is more accurate when a low number of representatives were selected.

Figure 5.16 show the measurement that performed highest in relation to what a random representative selection would result in. The clustering is only separating a few noise points and grouping the rest of the data into one cluster. This gives an indication like figures 5.14 and 5.15 that the initial cluster representation selections are the most accurate. This result shows that the initial representative selection has a high chance of being better than a random assignment would be.

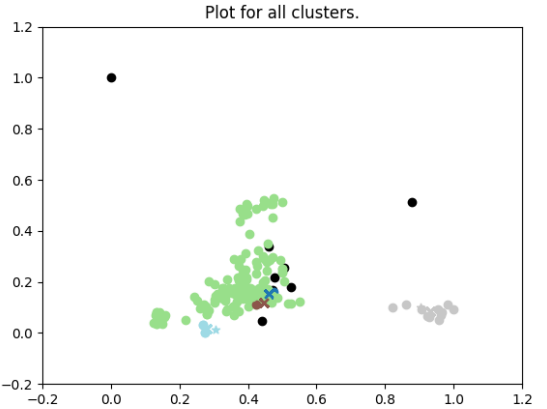

Figure 5.14 The figure shows a scatter plot of the DBSCAN clustering with the highest coverage in relation to optimal selection. The embedding was word2vec (Google news) reduced to 12 dimensions, with 5 clusters formed, represented by different colors. The noise points are black. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

Figure 5.15 The figure shows a scatter plot of the DBSCAN clustering with the highest coverage in relation to optimal selection. The embedding was spaCy reduced to 3 dimensions, with 5 clusters formed, represented by different colors. The noise points are black. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

Figure 5.16 The figure shows a scatter plot of the DBSCAN clustering with the highest coverage in relation to random selection. The embedding was word2vec reduced to 2 dimensions, with 1 cluster formed. The noise points are black. The stars show the data points that are cluster representatives, crosses show the centroids of clusters.

In summary, the results presented in the section shows that the DBSCAN approach overall achieves higher coverage results than the K-means approach. The only coverage result that was higher for the K-means approach was the coverage evaluation towards what optimal representative selection would result in. One factor that could skew the comparison is that the DBSCAN approach achieved the result with a higher number of clusters.

6

Discussion

In the discussion, we will discuss the possible ways that scenario selection using unsupervised learning could be improved and what promise it shows in the field of autonomous vehicles. In section 6.1, the way that we compared similarities is discussed and how better comparisons can be achieved. In section 6.2, we discuss the size of the dataset and possible problems associated with it. In section 6.3 we look at the clustering algorithms used in the methodology and what other clustering algorithms should be evaluated. In section 6.4, we discuss the usage of parameterization as a way to represent disengagements. In section 6.5 we discuss the evaluation metrics used to determine the performance of the model. Lastly, in section 6.5, we discuss the performance of the scenario selection methodology and unsupervised learning's viability to be used in the field.

6.1 Improving similarity comparisons

RQ1 concerns the ability to capture the similarities and differences of disengagement scenarios. The K-mean approach was able to reach a coverage of 41% of the parameter values for the disengagements and DBSCAN achieved a coverage of 53%. This reveals that the K-means approach fails to represent the majority of potential parameter values leading to disengagement, while the DBSCAN approach successfully captures slightly over half of them. In the background, we established that testing unexpected edge cases is central to autonomous vehicle testing. The approach presented in the thesis would leave many possible values untested. Therefore it can't be applied to the field in its current state. The low coverage suggests that there are issues with both the chosen embeddings for disengagement scenarios and the clustering approaches.

If the embeddings accurately represented the disengagements, we would expect to observe a high coverage result in measurements with effective clustering. However, the highest coverage results were not attained with the same embeddings as those that achieved high silhouette scores, low DBI, and a high CH index, which are indicators of effective clustering. The embeddings that achieved the highest scor-

ing in these metrics were embeddings with less spread and a high level of dimensionality reduction. If the embeddings effectively represented the similarity among disengagements, rather than just general language, the expectation is for them to have a high level of separation. However, the word2vec (Google News) embedding achieved the highest silhouette score and CH index, while also recording the lowest DBI score. The embeddings in the word2vec (Google news) model show a very tight clustering of data points in the plots, which is not ideal for separating different disengagement scenarios. The reason for the models poor separation could be that the word2vec (Google news) model is trained on a very large dataset. This means that the terms used in the disengagement scenarios may be seen as very similar in terms of a general comparison. Due to the specific application that we desire to use them in this is a problem. We wish to capture the small differences that separate disengagement scenarios rather than just being able to label them as autonomous vehicle related texts. With the K-means approach, both word2vec and doc2vec scored roughly equal or higher to the pre-trained models in terms of coverage. Considering that the training is performed on a very small dataset which is also the same used for assessment, the embeddings could be too well fitted with the data. But the result does warrant that the embeddings used for similarity comparison need to be worked on. It is also possible that the preprocessing could have stripped information necessary to capture the differences between disengagements. However, the preprocessing steps steps used in my work are commonly used in natural language processing tasks. Therefore, embeddings derived from a larger dataset within the field of autonomous vehicles is needed to evaluate the feasibility of a clustering approach of disengagements.

Another noteworthy finding is that dimensionality reduction appeared to significantly improve the CH index, DBI and silhouette score for all clustering approaches. PCA dimensionality reduction aims to retain the variance of data, but it does however filter out possible noise in the data. This could mean that the high level of dimensionality reduction leads to good clustering by reducing the variance due to noise in the data. In regards to value coverage, the reduction also seem to improve the clustering which may indicate that the embeddings can be subject to the curse of dimensionality. The dataset is very small for the project and the embeddings can capture many dimensions of noise that are not significant for separating disengagement scenarios.

6.2 The size of the dataset

One central challenge to the clustering approach is the small size of the dataset. The purpose of selecting cluster representatives is to reduce the number of required tests while retaining high coverage. However, the high variance of disengagements in the disengagement reports needs a high number of cluster representatives to achieve a high coverage. With the low coverage achieved using the proposed approach, the

clustering does not work at a high enough efficiency to be used in the field. The need for a high certainty in the safety of autonomous vehicles requires firstly that a testing approach achieves significant coverage and secondly a high efficiency.

6.3 Choosing an appropriate clustering algorithm

RQ2 concerns the clustering of disengagement scenarios with unsupervised learning. As discussed for RQ1, all clustering approaches achieved low coverage of the dataset. The evaluation highlights that it is possible to attain a high silhouette score and CH index while maintaining a low DBI for the chosen approaches. The results indicate that DBSCAN can have some advantages over the K-means approach. DBSCAN does not require a predetermined number of clusters for initialization. Its density-based approach also enhances its robustness in dealing with data points that may be difficult to assign to clusters. The approach's potential to find noise points could also improve the result. Some disengagements could be too different to be clustered with other scenarios within the limited dataset. K-means requires every data point to be assigned to a cluster even if the cluster quality is decreased. Edge cases could be grouped with dissimilar data points, hiding unique values if the wrong number of clusters are selected. In DBSCAN, the appropriate number of clusters can be based on setting a similarity radius (ε) and a minimum number of points (min. samples) to initiate clusters. The approach of determining the optimal number of clusters using the elbow method is based on stopping when the cluster quality is not improved by a higher number of clusters. The high cluster quality in terms of CH index, silhouette score and DBI, did not seem to be associated with a high coverage value. This means that the stopping criterion based on a cluster quality metric (distortion scoring) may be unsuited for the goal of clustering data of a very high variance. The reason that DBSCAN outperformed the K-means clustering may predominately be associated with a higher number of cluster representatives being selected for the DBSCAN approach, but in terms of determining the optimal number of clusters the DBSCAN may be favorable.

6.4 Representing disengagements with parameterization

RQ3 addresses representing scenarios in a standardized way. This approach will also be discussed. We used parameterization to represent disengagements in the thesis. The parameterization was performed manually by the author as a way to evaluate the clustering result. In the thesis it was used to evaluate clustering. In that regard the approach shows promise. The clustering would have needed to be interpreted manually for all scenarios to determine coverage if the dataset was not parameterized to reveal coverage results. The parameterization serves as an important tool to determine the correlation between cluster performance and effective coverage of parameter values. In terms of time required, the task is quite cumbersome and is

applied to all scenarios as a means of evaluation. In a real application, the only disengagements that would be parameterized is the cluster representations in order to create a test suite.

6.5 Assessing clustering performance and coverage

Silhouette score, DBI and CH index are some of the most common evaluation metrics for K-means clustering quality. In this thesis they were chosen to see if the clustering could accurately separate different disengagements. Though the metrics are commonly used, one drawback is that they can not be applied to the DBSCAN approach for evaluation. The metrics indicated that K-means could form well-defined clusters, ensuring that the embedded entries were being grouped well according to their numerical differences and similarities. However, as revealed previously in the discussion the cluster quality does not seem to correlate to high coverage of the dataset. This issue is most likely not only a problem caused by clustering but rather caused by the problems with embeddings previously discussed.

Moving on to the test coverage of the dataset, the evaluation approach can be discussed. The evaluation was based on the ratio of covered parameter values in relation to the total number of parameter values. An alternative approach would be to see how large portion of the disengagement scenarios could be tested with the representations. It can be argued that the chosen approach of evaluation does not reveal the actual coverage achieved of the dataset. However, in terms of industrial application, less common causes of disengagements are also of importance. The first choice provides a way to better see how large portion of possible values can be tested with the clustering. Whereas the assessment of scenario coverage would reveal more information in regards to if the most common parameters and their most common values have been accurately selected. This validates the choice of parameter value coverage as evaluation in the thesis.

Another choice of coverage that needs to be discussed is the evaluation of coverage in relation to what optimal representative selection would achieve. This method of evaluation was included together with the comparison to random selection. These metrics provide insight in how much higher performance can be achieved as well as a comparison to what level of coverage can be attributed to chance. The results shows that the initial cluster assignments are most accurate. This is indicated with the low number of clusters that the highest measurements were achieved with. However, it is worth nothing that the result may not be highly reliable. As this behavior is expected with the first assignments being less prone to selecting disengagements with overlapping parameter values, which may account for the result.

One choice that have been made in regards to the results presented is to only show the best scoring measurements. The reasoning for this evaluation process is that the main focus of the thesis was to determine if the approach could be used for cluster representative selection. As the highest performing results were not indicat-

ing that the approach was efficient, further evaluation of worst-case scenarios were not included. Such an evaluation could however be insightful if the approach could be optimized further.

6.6 Improving performance of the model

RQ4 concerns the coverage evaluation of the clustering approaches in the thesis. The evaluation was devised to test the approach in two stages. First, with silhouette score, CH index and DBI to determine if the approach formed well-defined clusters. Secondly, to evaluate the coverage of the dataset achieved by the complete approach.

One final question is worth discussing in regards to the coverage achieved with the thesis approach. That is, what level of improvements are necessary to make it usable in its field. The current selection of 19 representative scenarios only manages to achieve a coverage of approximately 41% of parameter values with K-means and 53% for 35 clusters with DBSCAN. Considering that the selection of more cluster representatives comes with diminishing returns due to parameter value overlap between scenarios, a coverage of 90% would require a high number of representative disengagement scenarios. This indicates that major improvements are needed in order for the unsupervised learning approach presented in this thesis to be applicable to the field.

7

Conclusion

For the research questions, we are now able to draw some conclusions. Our conclusion regarding RQ1 is that the similarity between disengagements can not be compared accurately using the approach applied in the thesis. This seems to be due to a lack of detail in the language processing tools applied for similarity comparisons. To enhance similarity comparisons, a field-oriented form of embedding needs to be evaluated which can better capture the differences between autonomous vehicles disengagements.

For RQ2 our conclusion is that the clustering approach presented is not ready for practical application without additional optimization. Future research should aim to evaluate the approach for larger datasets. Other clustering algorithm need to be evaluated in order to handle the large variety of disengagement scenarios. Finally, approaches that can determine the optimal number of clusters for data with high variety should be implemented in the approach.

For RQ3 the conclusion is that the approach shows promise. In terms of evaluation, it provides a way to determine if the clustering actually achieves a high level of coverage. This evaluation method also seems unprecedented and therefore could benefit substantially to the field. In terms of crafting a test suite, it provides the ability to combine the features of different test scenarios efficiently to test a large amount of non-exclusive parameters together.

For RQ4 our conclusion is that the clustering approach is unable to reach higher coverage than 53%. The evaluation process devised does however show some applicability in the field of testing clustering approaches for autonomous vehicles. The approach of cluster representation selection using unsupervised learning has proven to have many possible points of failure. The evaluation approach used with parameter value coverage can give a good insight into how well representation selection can cover the full dataset.

The main conclusion of the thesis is therefore that the clustering approach devised need more optimization to be useful. At its current stage, the achieved coverage is too low for practical implementation in the field of autonomous vehicle testing.

Though, the results show that the full approach can not currently be implemented. An area where the thesis can benefit the field is in the way that scenarios were parameterized and used in coverage evaluation. The representation of scenarios in a parameterized way has been utilized in earlier studies, their usage to evaluate coverage with representative scenario selection is unprecedented. This is a valuable addition as the project has shown that the clustering metrics (silhouette score, DBI and CH index) could not be used to find the optimal clustering in terms of coverage. The usage of parameterization to evaluate coverage provides an approach to better reveal coverage results. It is also an evaluation approach that can be applied to a larger range of clustering approaches compared to DBI, silhouette score and CH index.

We cannot conclusively determine if the proposed approach can be optimized to the level that the industry requires. However, some parts of the approach should be re-evaluated when better embeddings can be achieved and the data available has increased. My hope is that this thesis can serve as a base for future research. The focus of future research should be to find suitable embeddings for the context of autonomous vehicles. Larger datasets of autonomous vehicle disengagements are also crucial to reap the benefits that unsupervised learning can have in the field of autonomous vehicle testing.

Data availability

The data used in this paper was accessed using the California DMV homepage and by contacting the CA DMV using the following address AVarchive@dmv.ca.gov.

Bibliography

[1] J. C. Spence, Y.-B. Kim, C. G. Lamboglia, C. Lindeman, A. J. Mangan, A. P. McCurdy, J. A. Stearns, B. Wohlers, A. Sivak, and M. I. Clark, "Potential Impact of Autonomous Vehicles on Movement Behavior: A Scoping Review," en, *American Journal of Preventive Medicine*, **58**:no. 6, e191–e199, ISSN: 0749-3797. DOI: 10.1016/j.amepre.2020.01.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749379720300593 (visited on 2023-08-07).

[2] R. Hussain and S. Zeadally, "Autonomous cars: Research results, issues, and future challenges," *IEEE Communications Surveys & Tutorials*, **21**:no. 2, Conference Name: IEEE Communications Surveys & Tutorials, pp. 1275–1313, ISSN: 1553-877X. DOI: 10.1109/COMST.2018.2869360.

[3] E. Esenturk, D. Turley, A. Wallace, S. Khastgir, and P. Jennings, "A data mining approach for traffic accidents, pattern extraction and test scenario generation for autonomous vehicles," *International Journal of Transportation Science and Technology*, **1**:no. 1, pp. 1–10. DOI: 10.1016/j.ijtst.2022.10.002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2046043022000867.

[4] Z. Micskei, Z. Szatmári, J. Oláh, and I. Majzik, "A concept for testing robustness and safety of the context-aware behaviour of autonomous systems," in G. Jezic, M. Kusek, N.-T. Nguyen, *et al.* (Eds.), *Agent and Multi-Agent Systems. Technologies and Applications*, ser. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012, pp. 504–513, ISBN: 978-3-642-30947-2. DOI: 10.1007/978-3-642-30947-2_55.

[5] J. Cyriac, *Assessment and assurance of autonomous vehicle safety*, Thesis, University of Illinois at Urbana-Champaign, 2, 2020. [Online]. Available: https://hdl.handle.net/2142/109609 (visited on 2023-06-14).

[6] M. Martínez-Díaz and F. Soriguera, "Autonomous vehicles: Theoretical and practical challenges," *Transportation Research Procedia*, **33**, XIII Conference on Transport Engineering, CIT2018, pp. 275–282, ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2018.10.103. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146518302606.

[7] S. Liu and L. F. Capretz, "An analysis of testing scenarios for automated driving systems," in, *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, IEEE, Honolulu, HI, USA, 2021, pp. 622–629, ISBN: 978-1-72819-630-5. DOI: 10.1109/SANER50967.2021.00078. [Online]. Available: https://ieeexplore.ieee.org/document/9426042/ (visited on 2023-06-14).

[8] Q. Song, K. Tan, P. Runeson, and S. Persson, "Critical scenario identification for realistic testing of autonomous driving systems," *Software Quality Journal*, ISSN: 1573-1367. DOI: 10.1007/s11219-022-09604-2. [Online]. Available: https://doi.org/10.1007/s11219-022-09604-2 (visited on 2023-06-14).

[9] A. Erdogan, E. Kaplan, A. Leitner, and M. Nager, "Parametrized end-to-end scenario generation architecture for autonomous vehicles," in, *2018 6th International Conference on Control Engineering & Information Technology (CEIT)*, 2018, pp. 1–6. DOI: 10.1109/CEIT.2018.8751872.

[10] B. Gangopadhyay, S. Khastgir, S. Dey, P. Dasgupta, G. Montana, and P. Jennings, "Identification of test cases for automated driving systems using bayesian optimization," in, *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 1961–1967. DOI: 10.1109/ITSC.2019.8917103.

[11] E. de Gelder and J.-P. Paardekooper, "Assessment of automated driving systems using real-life scenarios," in, *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 589–594. DOI: 10.1109/IVS.2017.7995782.

[12] J. Tao, Y. Li, F. Wotawa, H. Felbinger, and M. Nica, "On the industrial application of combinatorial testing for autonomous driving functions," in, *2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 2019, pp. 234–240. DOI: 10.1109/ICSTW.2019.00058.

[13] S. Khastgir, S. Brewerton, J. Thomas, and P. Jennings, "Systems approach to creating test scenarios for automated driving systems," *Reliability Engineering & System Safety*, **215**, p. 107 610, ISSN: 09518320. DOI: 10.1016/j.ress.2021.107610. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0951832021001551 (visited on 2023-06-14).

[14] H. Alghodhaifi and S. Lakshmanan, "Autonomous Vehicle Evaluation: A Comprehensive Survey on Modeling and Simulation Approaches," *IEEE Access*, **9**, Conference Name: IEEE Access, pp. 151 531–151 566, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3125620.

[15] D. J. Fremont, E. Kim, Y. V. Pant, S. A. Seshia, A. Acharya, X. Bruso, P. Wells, S. Lemke, Q. Lu, and S. Mehta, "Formal Scenario-Based Testing of Autonomous Vehicles: From Simulation to the Real World," in, *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–8. DOI: 10.1109/ITSC45102.2020.9294368.

[16] California Department of Motor Vehicles, *DISENGAGEMENT REPORTS*, https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/disengagement-reports/, Accessed on 01-06-2023, 2023.

[17] A. Sinha, V. Vu, S. Chand, K. Wijayaratna, and V. Dixit, "A crash injury model involving autonomous vehicle: Investigating of crash and disengagement reports," *Sustainability*, **13**:no. 14, p. 7938, ISSN: 2071-1050. DOI: 10.3390/su13147938. [Online]. Available: https://www.mdpi.com/2071-1050/13/14/7938 (visited on 2023-06-14).

[18] C. Lv, D. Cao, Y. Zhao, D. J. Auger, M. Sullman, H. Wang, L. M. Dutka, L. Skrypchuk, and A. Mouzakitis, "Analysis of autopilot disengagements occurring during autonomous vehicle testing," *IEEE/CAA Journal of Automatica Sinica*, **5**:no. 1, pp. 58–68, ISSN: 2329-9266, 2329-9274. DOI: 10.1109/JAS.2017.7510745. [Online]. Available: https://ieeexplore.ieee.org/document/8232590/ (visited on 2023-06-14).

[19] SAE, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," Society of Automotive Engineers, Standard SAE J3016, 2021.

[20] International Organization for Standardization, *Road vehicles - functional safety, Part 1: Vocabulary*, Standard, 2018.

[21] P. Koopman, U. Ferrell, F. Fratrik, and M. Wagner, "A Safety Standard Approach for Fully Autonomous Vehicles," en, in A. Romanovsky, E. Troubitsyna, I. Gashi, *et al.* (Eds.), *Computer Safety, Reliability, and Security*, ser. Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 326–332, ISBN: 978-3-030-26250-1. DOI: 10.1007/978-3-030-26250-1_26.

[22] *Road vehicles – safety of the intended functionality*, ISO Standard, International Organization for Standardization, 2022.

[23] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE International Journal of Transportation Safety*, **4**:no. 1, pp. 15–24, ISSN: 2327-5634. DOI: 10.4271/2016-01-0128. [Online]. Available: https://www.sae.org/content/2016-01-0128/ (visited on 2023-06-14).

[24] T. F. Koné, E. Bonjour, E. Levrat, F. Mayer, and S. Géronimi, "Safety demonstration of autonomous vehicles: A review and future research questions," in G. A. Boy, A. Guegan, D. Krob, *et al.* (Eds.), *Complex Systems Design & Management*, Springer International Publishing, Cham, 2020, pp. 176–188, ISBN: 978-3-030-34843-4. DOI: 10.1007/978-3-030-34843-4_15.

[25] Y. Su and L. Wang, "Integrated framework for test and evaluation of autonomous vehicles," *Journal of Shanghai Jiaotong University (Science)*, **26**:no. 5, pp. 699–712, ISSN: 1995-8188. DOI: 10.1007/s12204-021-2360-y. [Online]. Available: https://doi.org/10.1007/s12204-021-2360-y (visited on 2023-06-15).

[26] J. Norden, M. O'Kelly, and A. Sinha, *Efficient black-box assessment of autonomous vehicle safety*, 5, 2020. DOI: 10.48550/arXiv.1912.03618. arXiv: 1912.03618[cs,stat]. [Online]. Available: http://arxiv.org/abs/1912.03618 (visited on 2023-06-15).

[27] F. Indaheng, E. Kim, K. Viswanadha, J. Shenoy, J. Kim, D. J. Fremont, and S. A. Seshia, *A scenario-based platform for testing autonomous vehicle behavior prediction models in simulation*, 13, 2021. DOI: 10.48550/arXiv.2110.14870. arXiv: 2110.14870[cs]. [Online]. Available: http://arxiv.org/abs/2110.14870 (visited on 2023-06-15).

[28] J. Bernhard, M. Schutera, and E. Sax, "Optimizing test-set diversity: Trajectory clustering for scenario-based testing of automated driving systems," in, *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 1371–1378. DOI: 10.1109/ITSC48978.2021.9564771.

[29] S. Xu, Z. Wang, L. Fan, X. Cai, H. Ji, S.-C. Khoo, and B. B. Gupta, "DeepSuite: A test suite optimizer for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, **23**:no. 7, pp. 9506–9517, ISSN: 1524-9050, 1558-0016. DOI: 10.1109/TITS.2021.3131808. [Online]. Available: https://ieeexplore.ieee.org/document/9646479/ (visited on 2023-06-14).

[30] F. Hauer, T. Schmidt, B. Holzmüller, and A. Pretschner, "Did we test all scenarios for automated and autonomous driving systems?" In, *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 2950–2955. DOI: 10.1109/ITSC.2019.8917326.

[31] C. Sippl, F. Bock, D. Wittmann, H. Altinger, and R. German, "From simulation data to test cases for fully automated driving and ADAS," in F. Wotawa, M. Nica, and N. Kushik (Eds.), *Testing Software and Systems*, ser. Lecture Notes in Computer Science, Springer International Publishing, Cham, 2016, pp. 191–206, ISBN: 978-3-319-47443-4. DOI: 10.1007/978-3-319-47443-4_12.

[32] Đ. Nalić, T. Mihalj, M. Baeumler, M. Lehmann, A. Eichberger, and S. Bernsteiner, *Scenario Based Testing of Automated Driving Systems: A Literature Survey*. 27, 2020. DOI: 10.46720/f2020-acm-096.

[33] J. Langner, H. Grolig, S. Otten, M. Holzäpfel, and E. Sax, "Logical scenario derivation by clustering dynamic-length-segments extracted from real-world-driving-data:" in, *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems*, SCITEPRESS - Science and Technology Publications, Heraklion, Crete, Greece, 2019, pp. 458–467, ISBN: 978-989-758-374-2. DOI: 10.5220/0007723304580467. [Online]. Available: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007723304580467 (visited on 2023-06-15).

[34] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for development, test and validation of automated vehicles," in, *2018 IEEE Intelligent Vehicles Symposium (IV)*, ISSN: 1931-0587, 2018, pp. 1821–1827. DOI: 10.1109/IVS.2018.8500406.

[35] G. Lou, Y. Deng, X. Zheng, M. Zhang, and T. Zhang, *Testing of autonomous driving systems: Where are we and where should we go?* 23, 2022. arXiv: 2106.12233[cs]. [Online]. Available: http://arxiv.org/abs/2106.12233 (visited on 2023-06-14).

[36] M. Wagner and P. Koopman, "A philosophy for developing trust in self-driving cars," in G. Meyer, S. Beiker (Eds.), *Road Vehicle Automation 2*, ser. Lecture Notes in Mobility, Springer International Publishing, Cham, 2015, pp. 163–171, ISBN: 978-3-319-19078-5. DOI: 10.1007/978-3-319-19078-5_14.

[37] H. Weber, J. Bock, J. Klimke, C. Roesener, J. Hiller, and R. Krajewski, "A framework for definition of logical scenarios for safety assurance of automated driving," English, *Traffic Injury Prevention*, **20**:no. sup1, Peer-Reviewed Journal for the 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV), S65–S70. DOI: 10.1080/15389588.2019.1630827. [Online]. Available: https://doi.org/10.1080/15389588.2019.1630827.

[38] H. Winner, K. Lemmer, T. Form, and J. Mazzega, "Pegasus—first steps for the safe introduction of automated driving," in G. Meyer, S. Beiker (Eds.), *Road Vehicle Automation 5*, ser. Lecture Notes in Mobility, Springer, Cham, 2019, pp. 185–195. DOI: 10.1007/978-3-319-94896-6_16. [Online]. Available: https://doi.org/10.1007/978-3-319-94896-6_16.

[39] Q. Goss, Y. AlRashidi, and M. İ. Akbaş, "Generation of modular and measurable validation scenarios for autonomous vehicles using accident data," in, *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 251–257. DOI: 10.1109/IV48863.2021.9575506.

[40] J. Bach, S. Otten, and E. Sax, "Model based scenario specification for development and test of automated driving functions," in, *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016, pp. 1149–1155. DOI: 10.1109/IVS.2016.7535534.

[41] F. Wotawa, "Testing autonomous and highly configurable systems: Challenges and feasible solutions," in D. Watzenig, M. Horn (Eds.), *Automated Driving: Safer and More Efficient Future Driving*, Springer International Publishing, Cham, 2017, pp. 519–532, ISBN: 978-3-319-31895-0. DOI: 10.1007/978-3-319-31895-0_22. [Online]. Available: https://doi.org/10.1007/978-3-319-31895-0_22 (visited on 2023-06-14).

[42] E. Esenturk, A. G. Wallace, S. Khastgir, and P. Jennings, "Identification of Traffic Accident Patterns via Cluster Analysis and Test Scenario Development for Autonomous Vehicles," *IEEE Access*, **10**, Conference Name: IEEE Access, pp. 6660–6675, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3140052.

[43] Y. Zhang, X. J. Yang, and F. Zhou, "Disengagement cause-and-effect relationships extraction using an nlp pipeline," *IEEE Transactions on Intelligent Transportation Systems*, **23**:no. 11, pp. 21 430–21 439. DOI: 10.1109/TITS.2022.3186248.

[44] K.-W. Wu, W.-F. Wu, C.-C. Liao, and W.-A. Lin, "Risk Assessment and Enhancement Suggestions for Automated Driving Systems through Examining Testing Collision and Disengagement Reports," en, *Journal of Advanced Transportation*, **2023**, A. Severino, (Ed.), pp. 1–18, ISSN: 2042-3195, 0197-6729. DOI: 10.1155/2023/3215817. [Online]. Available: https://www.hindawi.com/journals/jat/2023/3215817/ (visited on 2023-06-14).

[45] F. M. Favarò, S. O. Eurich, and N. Nader, "Analysis of disengagements in autonomous vehicle technology," in, *2018 Annual Reliability and Maintainability Symposium (RAMS)*, ISSN: 2577-0993, 2018, pp. 1–7. DOI: 10.1109/RAM.2018.8463123.

[46] F. M. Favarò, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju, "Examining accident reports involving autonomous vehicles in California," en, *PLOS ONE*, **12**:no. 9, X. Hu, (Ed.), e0184952, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0184952. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0184952 (visited on 2023-06-14).

[47] Q. Lu, J. G. Conrad, K. Al-Kofahi, and W. Keenan, "Legal document clustering with built-in topic segmentation," in, *Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM, Glasgow Scotland, UK, 24, 2011, pp. 383–392, ISBN: 978-1-4503-0717-8. DOI: 10.1145/2063576.2063636. [Online]. Available: https://dl.acm.org/doi/10.1145/2063576.2063636 (visited on 2023-06-14).

[48] W.-N. Zhang, T. Liu, Y. Yang, L. Cao, Y. Zhang, and R. Ji, "A Topic Clustering Approach to Finding Similar Questions from Large Question and Answer Archives," en, *PLoS ONE*, **9**:no. 3, D. Abbott, (Ed.), e71511, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0071511. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0071511 (visited on 2023-06-14).

[49] B. B. Kadayat and E. Eika, "Impact of Sentence Length on the Readability of Web for Screen Reader Users," en, in M. Antona, C. Stephanidis (Eds.), *Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies*, ser. Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 261–271, ISBN: 978-3-030-49282-3. DOI: 10.1007/978-3-030-49282-3_18.

[50] V. Mohan, "Preprocessing Techniques for Text Mining - An Overview."

[51] M. Honnibal and I. Montani, *Spacy: Industrial-strength natural language processing in python*, 2021. [Online]. Available: https://spacy.io/.

[52] R. Pramana, Debora, J. J. Subroto, A. A. S. Gunawan, and Anderies, "Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity," in, *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, 2022, pp. 1–6. DOI: 10.1109/ICITDA55840.2022.9971451.

[53] B. Srinivasa-Desikan, *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd, 2018.

[54] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in, *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013. [Online]. Available: https://openreview.net/forum?id=r37lOXWCe.

[55] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in, *International Conference on Machine Learning (ICML)*, 2014, pp. 1188–1196. [Online]. Available: http://proceedings.mlr.press/v32/le14.html.

[56] S. Tahvili, L. Hatvani, M. Felderer, W. Afzal, and M. Bohlin, "Automated functional dependency detection between test cases using doc2vec and clustering," in, *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 2019, pp. 19–26. DOI: 10.1109/AITest.2019.00-13.

[57] G. Di Gennaro, A. Buonanno, and F. A. N. Palmieri, "Considerations about learning Word2Vec," en, *The Journal of Supercomputing*, **77**:no. 11, pp. 12 320–12 335, ISSN: 1573-0484. DOI: 10.1007/s11227-021-03743-2. [Online]. Available: https://doi.org/10.1007/s11227-021-03743-2 (visited on 2023-08-29).

[58] X. Rong, *Word2vec Parameter Learning Explained*, arXiv:1411.2738 [cs], 2016. [Online]. Available: http://arxiv.org/abs/1411.2738 (visited on 2023-08-29).

[59] *Word2vec: Tool for computing continuous distributed representations of words*, [https://code.google.com/archive/p/word2vec/], Retrieved [2023-08-29].

[60] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in, *International Conference on Machine Learning (ICML)*, vol. 14, 2014, pp. 1188–1196.

[61] S. JUGRAN, A. KUMAR, B. S. TYAGI, and V. ANAND, "Extractive Automatic Text Summarization using SpaCy in Python & NLP," in, *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 582–585. DOI: 10.1109/ICACITE51222.2021.9404712.

[62] L. H. Patil and M. Atique, "A novel approach for feature selection method TF-IDF in document clustering," in, *2013 3rd IEEE International Advance Computing Conference (IACC)*, 2013, pp. 858–862. DOI: 10.1109/IAdCC.2013.6514339.

[63] D. Rani, R. Kumar, and N. Chauhan, "Study and Comparision of Vectorization Techniques Used in Text Classification," in, *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2022, pp. 1–6. DOI: 10.1109/ICCCNT54827.2022.9984608.

[64] N. Tomašev and M. Radovanović, "Clustering Evaluation in High-Dimensional Data," en, in M. E. Celebi, K. Aydin (Eds.), *Unsupervised Learning Algorithms*, Springer International Publishing, Cham, 2016, pp. 71–107, ISBN: 978-3-319-24211-8. DOI: 10.1007/978-3-319-24211-8_4. [Online]. Available: https://doi.org/10.1007/978-3-319-24211-8_4 (visited on 2023-08-30).

[65] W. K. Vong, A. T. Hendrickson, D. J. Navarro, and A. Perfors, "Do Additional Features Help or Hurt Category Learning? The Curse of Dimensionality in Human Learners," en, *Cognitive Science*, **43**:no. 3, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12724, e12724, ISSN: 1551-6709. DOI: 10.1111/cogs.12724. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12724 (visited on 2023-09-26).

[66] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, **8**, Conference Name: IEEE Access, pp. 54 776–54 788, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2980942.

[67] R. Singh and S. Singh, "Text Similarity Measures in News Articles by Vector Space Model Using NLP," en, *Journal of The Institution of Engineers (India): Series B*, **102**:no. 2, pp. 329–338, ISSN: 2250-2114. DOI: 10.1007/s40031-020-00501-5. [Online]. Available: https://doi.org/10.1007/s40031-020-00501-5 (visited on 2023-08-29).

[68] S. Polamuri, *Five most popular similarity measures implementation in python*, Dataaspirant, 2015. [Online]. Available: https://dataaspirant.com/2015/04/11/five-most-popular-similarity-measures-implementation-in-python/.

[69] J. Farjo, W. Masri, and H. Hajj, "Isolating failing test cases: A comparative experimental study of clustering techniques," in, *2013 Third International Conference on Communications and Information Technology (ICCIT)*, 2013, pp. 73–77. DOI: 10.1109/ICCITechnology.2013.6579525.

[70] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, **8**, Conference Name: IEEE Access, pp. 80 716–80 727, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2988796.

[71] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," *Electrical Engineering and Computer Science - All Scholarship*. [Online]. Available: https://surface.syr.edu/eecs/43.

[72] S. Na, L. Xumin, and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," in, *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63–67. DOI: 10.1109/IITSI.2010.74.

[73] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: Past, present and future," in, *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 2014, pp. 232–238. DOI: 10.1109/ICADIWT.2014.6814687.

[74] T. M. Kodinariya and et al., "Review on determining number of cluster in k-means clustering," *International Journal*, **1.6**, pp. 90–95.

[75] Y. Developers. "Elbow method in yellowbrick," The scikit-yb developers. (2016-2019), [Online]. Available: https://www.scikit-yb.org/en/latest/api/cluster/elbow.html.

[76] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior," in, *2011 31st International Conference on Distributed Computing Systems Workshops*, ISSN: 2332-5666, 2011, pp. 166–171. DOI: 10.1109/ICDCSW.2011.20.

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, **12**, pp. 2825–2830.

[78] B. Bengfort, R. Bilbro, N. Danielsen, L. Gray, K. McIntyre, P. Roman, Z. Poh, *et al.*, *Yellowbrick*, version 0.9.1, 14, 2018. DOI: 10.5281/zenodo.1206264. [Online]. Available: http://www.scikit-yb.org/en/latest/.

[79] G. Chen, X. Ma, D. Yang, S. Tang, M. Shuai, and K. Xie, "Efficient approaches for summarizing subspace clusters into k representatives," en, *Soft Computing*, **15**:no. 5, pp. 845–853, ISSN: 1433-7479. DOI: 10.1007/s00500-010-0552-8. [Online]. Available: https://doi.org/10.1007/s00500-010-0552-8 (visited on 2023-09-01).

[80] N. Tomašev and M. Radovanović, "Clustering Evaluation in High-Dimensional Data," en, in M. E. Celebi, K. Aydin (Eds.), *Unsupervised Learning Algorithms*, Springer International Publishing, Cham, 2016, pp. 71–107, ISBN: 978-3-319-24211-8. DOI: 10.1007/978-3-319-24211-8_4. [Online]. Available: https://doi.org/10.1007/978-3-319-24211-8_4 (visited on 2023-08-30).

[81] P. G. Sapna and H. Mohanty, "Clustering test cases to achieve effective test selection," in, *Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India*, ser. A2CWiC '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 1–8, ISBN: 978-1-4503-0194-7. DOI: 10.1145/1858378.1858393. [Online]. Available: https://dl.acm.org/doi/10.1145/1858378.1858393 (visited on 2023-08-30).

[82] A. Dudek, "Silhouette Index as Clustering Evaluation Tool," en, in K. Jajuga, J. Batóg, and M. Walesiak (Eds.), *Classification and Data Analysis*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, Springer International Publishing, Cham, 2020, pp. 19–33, ISBN: 978-3-030-52348-0. DOI: 10.1007/978-3-030-52348-0_2.

[83] J. Baarsch and M. E. Celebi, "Investigation of internal validity measures for k-means clustering," in, *Proceedings of the International Multiconference of Engineers and Computer Scientists*, Newswood Limited, 2012, pp. 14–16.

[84] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," in, *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 306–310. DOI: 10.1109/ICCMC48092.2020.ICCMC-00057.

[85] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, **9**:no. 3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2023-11-27

EXAMENSARBETE Unsupervised Learning-Based Test Scenario Selection using Autonomous Vehicle Disengagements
STUDENT Oskar Andersson
HANDLEDARE Qunying Song (LTH)
EXAMINATOR Emelie Engström (LTH)

Urval av testscenarier för förarlösa bilar

POPULÄRVETENSKAPLIG SAMMANFATTNING **Oskar Andersson**

Förarlösa bilar börjar nå offentliga vägar runt om i världen och många känner oro för vilka säkerhetsrisker det kan innebära. För att förbättra vår förmåga att testa förarlösa bilar har mitt examensarbete undersökt möjligheten att gruppera trafiksituationer baserat på likheter.

Mitt examensarbete har fokuserat på att effektivisera testning av förarlösa bilar. I ett första steg har jag grupperat olika trafiksituationer som förarlösa bilar har problem med att hantera. Genom att identifiera grupper av liknande situationer så kan vi minska mängden test som behöver genomföras då vi kan testa ett fall per grupp istället för varje enskilt fall om gruppernas scenarier är väldigt lika.

Förarlösa bilar har stor potential att öka trafiksäkerheten på lång sikt då många olyckor sker på grund av förares misstag. Vi måste dock först bevisa att förarlösa bilar kan tolka och hantera alla möjliga trafiksituationer. Processen att testa dem på offentliga vägar är både kostsam och kan leda till att stora resurser läggs på att upprepa scenarier som bilen redan hanterat. För att försäkra sig om att även ovanliga fall testas måste vi använda testresurser effektivt.

Att granska tusentals trafiksituationer och bedöma deras likheter är både svårt och tidskrävande. Jag har därför tränat en dator att göra bedömningarna genom maskininlärning. Datorn kan klara av uppgiften genom att vi först anger vilka egenskaper som är viktiga hos beskrivningarna. Därefter kan datorn skapa grupper av beskrivningarna baserat på vilken grad av likhet vi vill att beskrivningar i grupper ska uppnå.

Mitt projekt visade att metoden i nuläget inte kan användas för testning då den kan leda till att vi missar många testfall. För att förbättra metoden måste vidare undersökningar göras om vilka egenskaper hos beskrivningarna som ska avgöra deras likheter.

Att göra testning mer effektivt gör att vi kan säkerställa att den testning som genomförs faktiskt ställer nya krav på bilen och därmed ökar dess säkerhet. Det går exempelvis inte att säga att en förarlös bil är säker bara för att den har kört tusentals mil men aldrig mött en poliskontroll.

Allteftersom fler förarlösa bilar sätter sina hjul på offentliga vägar kommer inrapporteringen av situationer som skapar problem för bilarna öka. Genom att identifiera grupper av liknande scenarier med maskininlärning kan vi avgöra vilka unika situationer som behöver testas automatiskt. Framtiden med självkörande bilar kan bidra med en stor mängd fördelar så länge vi bekräftar att bilarna klarar av det oväntade!