# Behavioral Topic Modeling on Naturalistic Driving Data

Sebastiaan Merino
ANR: 196813

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Data Science: Business and Governance,
at the School of Humanities and Digital Sciences
of Tilburg University

Thesis committee:

Dr. M. Atzmueller
Dr. A.T. Hendrickson

Tilburg University
School of Humanities and Digital Sciences
Department of Communication and Information Sciences
Tilburg center for Cognition and Communication (TiCC)
Tilburg, The Netherlands
December, 2018

# Abstract

Identifying risky driving behavior is of central importance for increasing traffic safety. This research tackles the task of analyzing naturalistic driving data captured by interpretable data science methods using in-vehicle sensors. In particular, this thesis focuses on symbolic time-series abstraction and the subsequent behavioral profile identification using Probabilistic Topic Modeling (PTM). Originally, PTM is applied in text-based studies to uncover latencies in text. By applying symbolic time-series abstraction methods on driving data, this study was able to present rides as documents and occurrences (i.e., speed and acceleration) as words. Twenty-four real-life rides presented 24 documents, and were analyzed using Latent Dirichlet Allocation (LDA). By applying different strategies in topic modeling, a variety of two-, to four-topic models were detected with meaningful behavioral topics. The evaluation of these strategies indicates that inclusion of bi-, and trigrams leads to more interpretable topics. Topics that were extracted from the driving data varied from city driving to highway driving, with in between nuances such as driving during rush hour and variation in acceleration due to obstacles on the road (i.e., traffic lights). To compare alert versus unalert participants, a Psychomotor Vigilance Test (PVT) was included in each experiment, which made it possible to objectively label participants. The results indicate that in general the probability of topic distribution was equal for alert and unalert participants. Future research should benefit applying the methodology applied in this study and acquire larger datasets as this would lead to a larger variety of topics. Besides that, inclusion of PVT as a long-term study would benefit to objectively control for alertness. Last, techniques, which improve symbolic time-series abstraction, should be applied to be able to translate naturalistic driving data into more meaningful words.

# Acknowledgements

First, I would like to express my very great appreciation to my thesis supervisor Dr. Martin Atzmueller of the Data Science: Business and Governance Master at Tilburg University. Dr. Atzmueller was very helpful and throughout the process of writing my master thesis, his constructive advise and suggestions helped me through the development of my work. His willingness to give his time so generously has been very much appreciated.

Also, I would like to thank Inge Wirken from Crossyn Automotive B.V. for offering me the opportunity to write my thesis at Crossyn. The usage of the hardware of Crossyn and the dataset made it possible to develop myself more into data science.

I would also like to thank Yannick Colastica from Toyota Netherlands for providing me a vehicle to conduct my experiments. It was a great experience for me to conduct the experiments in this safe and modern car and I would like to thank Toyota Netherlands for this tremendous gesture of providing a car.

Last, I would like to thank my family and friends for all the moral support and advise they gave throughout the development of my thesis. In particular, I would like to thank my girlfriend Michaela for being such a great support during my thesis. Michaela, thank you so much for all encouraging words during my process and for being there.

# Contents

# 1 Introduction

## 1.1 Motivation

An increase in road accidents has become a central issue for researchers to predetermine risky driving behavior, which increases the chance of road fatalities. The Dutch government has the objective to keep the number of road fatalities by 2020 under 500 per year (Aarts, Weijermars, Schoon, & Wesemann, 2008). However, the Dutch "Smart Traffic Accident Reporting" (STAR-initiative) argued that the goal for 2020 is unreachable as road fatalities keep increasing. More specifically, 597 road fatalities have been recorded in 2016 in the Netherlands. A gradual increase of 613 road fatalities was recorded in 2017 (CBS, 2018). In order to provide more detailed insights into real-life driving behavior, sensor-based data science is an important emerging research area, i.e., for diminishing the trend in road accidents, and to enhance overall traffic safety.

## 1.2 Background

Extensive research has already provided many insights in the field of road safety. For example, Zhang, Yau, Zhang, and Li (2016) mentioned in their study that an increase in motorization has lead to "severe traffic-related causalities" (Zhang et al., 2016, p. 34). In January 2018, the amount of vehicles in the Netherlands increased with 2% to 12.5 million vehicles. Yang, Li, Guan, Zhang, and Fan (2018), investigated whether high-density in traffic had an influence on driving behavior and concluded that lane changing and overtaking cars led to a significant increase. Also Cantin, Lavallière, Simoneau, and Teasdale (2009) mentioned lane changing and overtaking cars has negative effects on traffic safety. Thus, an increase of vehicles on the Dutch road has a negative influence on traffic safety.

Besides traffic density having an influence on traffic safety, research has also been conducted to study driving behavior (Bener, Lajunen, Özkan, Yildirim, & Jadaan, 2017; H.-Y. W. Chen, Donmez, Hoekstra-Atwood, & Marulanda, 2016; Garbarino et al., 2017). Studies included self-reports of large populations to study the behavior of participants while driving a car. The focus in the studies was to examine driving behavior by analyzing risks that have a negative influence such as, drowsiness, intoxication, aggression, or distraction. While it is indisputable that well designed questionnaires allow researchers to test hypotheses, Wohleber and Matthews (2016) con-

cluded that overconfidence might lead to implications in results of studies which utilize questionnaires to determine driver safety. A reason for this is the "belief that one possesses a greater competence than one's peers" (Wohleber & Matthews, 2016, p. 265). Overestimation, has been known to cause the illusion of control where an individual, who is in an adverse state of driving, might perceive himself in a controllable state. Studies revealed that while individuals perceived themselves in a controllable state, cognitive tasks were performed significantly worse. Thus, overestimation leads to underestimation of risks, which leads individuals to avoid precautions such as resting, hands-free driving, or not making use of a mobile phone while driving (Saxby, Matthews, & Neubauer, 2017).

Another disadvantage in studies, which studied risky driving behavior, is the quality of pre-crash information. More specifically, examples were given by Neale, Dingus, Klauer, Sudweeks, and Goodman (2005) in which police reports failed to mention the factual cause of an accident. Instead, reports were limited to determine the car as a rear-end collision and that the cause of it were cars following too close to another. It would have been more valuable to learn the state of the driver to determine the risk factors of incautious drivers. Therefore, attempts have been made to determine driving behavior from a drivers' point of view. For example, Radun, Radun, Wahde, Watling, and Kecklund (2015) were able to determine the risks factors and risk groups among truck drivers through self-reports, and concluded that truck drivers were aware of their fatigue state even before they started their shift. Such expositions as Radun et al. (2015) mentioned are of essence to evaluate risk groups and risk factors of driving behavior in order to educate and create awareness among drivers. However, experimental research techniques are available to predetermine potential adverse driving behavior and would therefore be more effective than post-accidental risk assessments.

As sensor-based data science emerged, state-of-the-art tools have made it possible to conduct more in-depth research to study real-life driving behavior. Recent studies have already been conducted in which driving behavior was analyzed by applying, mobile phones, stereo vision systems, GPS-sensors, wearable cameras, and heart rate monitors (Abouelnaga, Eraqi, & Moustafa, 2017; Battiato, Farinella, Gallo, & Giudice, 2018; Belakhdar, Kaaniche, Djemal, & Ouni, 2018; Botzer, Musicant, & Perry, 2017; Chowdhury, Chakravarty, Ghose, Banerjee, & Balamuralidhar, 2018; Park, Lee, Park, Seong, & Youn, 2018). Botzer et al. (2017) studied the effect of a smartphone collision warning application system (CWA) to condition potential risky driving behavior of participants. Battiato et al. (2018) applied stereo vision systems in their study to create a pedestrian detection system by applying a Traffic Conflicts Technique (TCT), which was able to detect unpredictable situations in traffic by measuring the interactions be-

tween pedestrians and buses. By analyzing spatial and temporal variables, risk impact levels could be created on intersections which provided more insights in traffic safety. Chowdhury et al. (2018) were able to identify drivers with high accuracy (82.3%) using solely GPS data. Wearable cameras were applied in the research of Abouelnaga et al. (2017), and by labelling postures of drivers, movements were classified using a deep neural network. Park et al. (2018) measured heart rate variability using an electrocardiogram (ECG)-signal to analyze low and high frequency features in heart rate, and were able to predict drowsiness by 93.11% using a Support Vector Machine-classifier. The advantage of theses studies is the use of naturalistic driving behavior (NDB). Unlike studies, which include self-reports to distinguish driving behavior, NDB obtains data from real-life situations. Besides that, sensor-based studies increase the ecological validity of NDB-research as non-intrusive sensors can be applied while obtaining experimental data. Moreover, data of multiple participants can be retrieved automatically as sensors can be installed simultaneously in vehicles.

## 1.3 Relevance

Although sensor-based apparatus have offered high quality data, studies are challenged. More specifically, McLaurin et al. (2014) mentioned NDB-analysis is challenged as "the large amount of data commonly collected during naturalistic driving studies makes comprehensive analysis prohibitive without some type of data reduction" (McLaurin et al., 2014, p. 2107). The authors further emphasize that data reduction should be conducted with precaution, as it might lead to omission of relevant data, or deformation of data structures. To cope with large time series data sets, McLaurin et al. (2014) included a time-series abstraction method, also referred as Symbolic Aggregate Approximation (SAX). Subsequently, Natural Language Processing (NLP) was applied which allowed to analyze the time series data with an unsupervised learning technique, referred as Probabilistic Topic Modelling (PTM). Originally, PTM has been applied in text mining studies as it analyzes text in documents. The output of PTM is a topic model which assigns occurrence-probabilities in text to a set of words. The probability distribution of these set of words are then assigned to as topics. As documents may contain multiple topics, PTM allows to classify topics in documents. This method has successfully been applied in time series data sets, as it has the potential to analyze large data sets while producing a comprehensive set of topics, which may reveal patterns that manual interference is not able to discover. McLaurin et al. (2014) were even able to identify drivers with Obstructive Sleep Apnea (OSA) compared to non-OSA-participants. A key issue which they have not treated in their study was to optimize

PTM by exploring optimization steps, such as "alternative word definitions, optimization of parameters, models that include $n$-grams, or patterns of multiple words" (McLaurin et al., 2014, p 2111). Therefore, this study explores steps which could potentially optimize PTM using naturalistic driving data (NDD).

## 1.4 Research question

Previously, McLaurin et al. (2014) suggested to further study optimization steps and to conclude whether clusters are found which lead to new insights in naturalistic driving behavior. Currently, Aksan, Dawson, Tippin, Lee, and Rizzo (2015) have exclusively followed up this previous study and applied alternative word definitions. However, their study lacked focus on optimization steps such as, optimizing parameters, using $n$-grams, and patterns of words. Besides that, different variables (i.e., steering wheel angle, and rate of change of steering wheel angle) were applied. This study will extend the study of McLaurin et al. (2014) by optimizing clusters. McLaurin et al. (2014) included lateral-, and longitudinal acceleration, and speed data in their symbolic representation of time series data. In the current study, a different in-vehicle sensor was applied which recorded GPS, linear acceleration and speed data. Thus, a distinctive symbolic representation will be included in the analysis (i.e., linear acceleration and speed). By applying this symbolic representation of time series data, this study is interested in finding patterns of naturalistic driving behavior. Thus, the main research question is defined as:

> *RQ 1: Which behavioral driving topics can be distinguished from naturalistic driving behavior?*

Before applying PTM in their experiment, McLaurin et al. (2014) mentioned that a complete corpus was included to the PTM-analysis, as they considered full inclusion of essence to ensure no relevant data was removed. However, this study hypothesizes that a combination of inclusion and exclusion is of essence in order to determine both general patterns and more subtile patterns. More specifically, pre-processing steps which include or exclude the importance of frequent or minimal occurring words need to be taken into account, as frequent occurring words might ignore more subtile differences in text González, Romero, Guerrero, and Calderón (2015). Therefore, a sub research question can be formulated as:

> *RQ 2: Which general and subtile behavioral driving topics can be distinguished from naturalistic driving behavior?*

McLaurin et al. (2014) controlled their experiments by collecting NDD of 26 participants (13 OSA patients and 13 non-OSA patients). Differences

in driving behavior were measured by comparing topic probabilities of both groups, and to measure significant differences. Unlike the aforementioned study, the current study had no access to a human subject pool with OSA-patients. Instead, this study is interested in detection of the alertness state of participants. To achieve this, a Psychomotor Vigilance Test (PVT) was included in each experiment. PVT is a widely applied test in studies to measure alertness, and reaction time of individuals to determine their alert state, and thus, adequacy of mental performance (Aryal, Ghahramani, & Becerik-Gerber, 2017). Analyzing psychomotor vigilance of participants had the objective to determine the alertness state of each participant, and to analyze whether alertness influences driving behavior. Thus, a second sub research question is formulated as:

> *RQ 3: Which differences in naturalistic driving behavior prevail between alert and unalert participants?*

NDD for this study were collected by conducting twenty-four real-life experiments. Crossyn provided the data set which was collected by using one in-vehicle GPS sensor, which was installed in a 2017 Toyota CH-R. GPS recordings determined speed, and linear acceleration. After data is preprocessed, corresponding steps will be conducted for PTM. Then, in order to establish a stronger PTM, optimization steps in NLP will be applied. Participant alertness was measured with a PVT, which was conducted on an iPad Pro 9.7 inch. The results of this experiment should provide insights, which enable answering the main research question, and sub research questions.

## 1.5 Significance

The study of McLaurin et al. (2014) constructed a framework for PTM, which applied NDD. Optimization in PTM benefits future research as it could potentially establish a more extensive framework for studies which would like to include unsupervised learning techniques. At the same time, the findings of this study should also contribute to future research which involves supervised learning techniques as the PTM-methodology has the potential to define differences in driving behavior of groups of people. Furthermore, this study aims to improve the predominant PTM-model, which contributes to the exploration of clusters which contain subtile distinctions compared to more general clusters.

In collaboration with Crossyn Automotive B.V.[1] in the Netherlands, the inclusion of an in-vehicle sensor, which recorded rides during experiments. Crossyn is a provider of services for car- and fleet-owners with the objective to improve the driving experience. Key aspects, which aim to improve this experience are, predictive maintenance, economic driving, and safe driving

[1] https://www.crossyn.com/crossyn

behavior. This study aims to provide new insights in safe driving behavior, which contributes to improvement of services of Crossyn. Furthermore, by creating strong algorithms in driving behavior, future handheld and in-vehicle applications can be created which act as state-of-the-art warning systems, which aim at reducing risky driving behavior, and thus, road fatalities.

## 1.6 Thesis outline

The remaining part of the paper has been divided into eight chapters.

- First, Chapter 2 provides an overview of the relevant literature. Section 2.1 briefly clarifies the background of machine learning algorithms, and Section 2.2 related work regarding cluster algorithms. Then, Section 2.3 describes steps which were defined in previous studies to convert time-series data to symbolic representation by applying SAX. Section 2.4 provides in-to depth knowledge of PTM by explaining its algorithm. Finally, Section 2.5 describes the background of PVT regarding its utility in research.

- Chapter 3 details the applied methods of this study. In Section 3.1, a detailed overview is given of steps that were conducted for data collection. Section 3.2 provides an overview of the data set and the preprocessing steps which were applied in order to prepare the data for PTM. Section 3.3 explains the steps for PTM, and provides an overview of the methodology for parameter selection. Finally, Section 3.4 is a detailed overview of preprocessing steps for PTM, which were necessary to answer the second research question.

- Chapter 4 analyses the results of the all experiments conducted for this study. Interpretation of the topics, in each topic model are described in detail, and visual representation is joined. In Section 4.2, results of the first experimental setup are presented, which are needed to answer the first research question. Second, Section 4.3 relates to results of the second part of experiments including different topic model strategies.

- Chapter 5 describes the experimental setup of the Psychomotor Vigilance Test, which was included in the driving experiments. In Section 5.1 the setup of the PVT, including procedure steps are described. Subsequently, steps are included in Section 5.3, which explain how the acquired results of the PVT are translated to label participants. In Sections 5.3.1 and 5.3.2 it is explained how labeled data is applied to evaluate the experiments of this study.

- Chapter 6 presents the results of the topic models, compared to the PVT results. In each section of this chapter, the results of each topic model strategy are described.
- Chapter 7 evaluates the results of all experiments and relates them to the research questions. In Section 7.1 a critical overview of the limitations in this study are provided. Then, Section 7.2 provides suggestions for future research.
- Finally, Chapter 8 concludes the results achieved in this study, and how the results have contributed to current and future research.

# 2 Related Work

Extensive research has already provided many insights in the field of road safety. The current study is related to previous established work in the field of behavioral topic modeling. In section 2.1, a brief overview is provided to conceptualize machine learning strategies. Subsequently, Section 2.2 distinguishes different types of cluster strategies. Third, section 2.3 describes how driving data can be symbolically represented by Symbolic Aggregate Approximation. Then, Topic Modeling is explained in section 2.4 and provides theory about the approach of this algorithm to large datasets. Finally, in Section 2.5 Psychomotor Vigilance Test is defined and discussed in detail on how it will be applied in this study.

## 2.1 Machine Learning

The field of machine learning distinguishes two types of approaches when analyzing data. Figure 2.1 provides an overview in which these approaches are divided into unsupervised and supervised learning techniques. Using this first approach, researchers have been able to analyze data in which "elements are clustered into different groups based on similarities and dissimilarities" (Virdi & Madan, 2018, p. 2). Unlike supervised learning techniques, in which data points are labelled so machine learning algorithms can classify them correspondingly, unsupervised learning techniques operate in the absence of labels. Recent studies, which applied naturalistic driving data to cluster their data, emphasized the importance of clustering techniques. For example, Li, Wang, Mo, and Zhao (2018) emphasized clustering data is of essence in naturalistic driving datasets as manually labelling these is very time consuming. Moreover, manual labelling of data is risky as underlying information in datasets may be missed while tracking real driving scenario's (Li et al., 2018). Thus, it is crucial to include unsupervised learning techniques in analyzing naturalistic driving data, in order to benefit time-efficiency and no underlying information is lost during analyses.

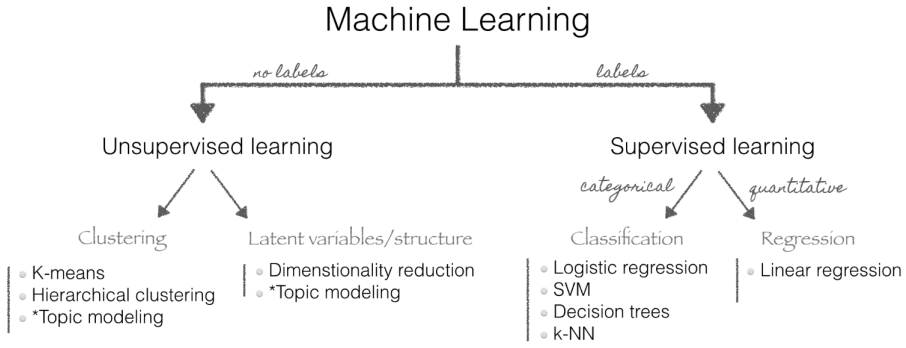



**Figure 2.1** Overview of Machine Learning applications divided into a spectrum of unsupervised and supervised learning. Adapted from the PyTexas conference 2015 in Texas, September 25, 2015. Retrieved September 4, 2018, from chdoig.github.io/pytexas2015-topic-modeling/#/2/1. Reprinted with permission. Image is licensed under CC BY 2.0 (Doig, 2015).

In the unsupervised learning spectrum, three types of clustering techniques can be distinguished, which are: hard-, hierarchical-, and soft/fuzzy clustering. Figure 2.2 provides a visual representation of these different types of cluster techniques. The vast majority of studies on clustering algorithms, based their analyses on hard clustering. More specifically, K-means clustering, which falls under the scope of hard cluster algorithms, has served the majority of researchers, who applied clustering analysis in their studies (Jain, 2010). K-means clustering, segments groups of data points into clusters by applying a squared error criterion (Jain, Murty, & Flynn, 1999). As hard clustering distinguishes groups of data, and forms homogeneous clusters, this technique has allowed studies to explore unlabelled data. For example, Yang, Ma, Zhang, Guan, and Jiang (2018) clustered two types of drivers in their study, which were: aggressive and non-aggressive drivers. Their data consisted of EEG-signals from participants who took part in a simulation study. Subsequently, after determining the clusters, they were able to apply a Support Vector Machine (SVM) algorithm as the clusters they had created served as labels. Yang, Ma, et al. (2018) applied k-Means clustering in their study, which in this case, is a typical form of hard clustering as participants were either labelled aggressive or unaggressive-, and stable or unstable behavior. The left part of figure 2.2, provides a visual representation of hard clustering. Here, the result of k-Means clustering is shown as clear distinctions are made between clusters, allowing researchers to classify each cluster as a label and apply supervised learning techniques.

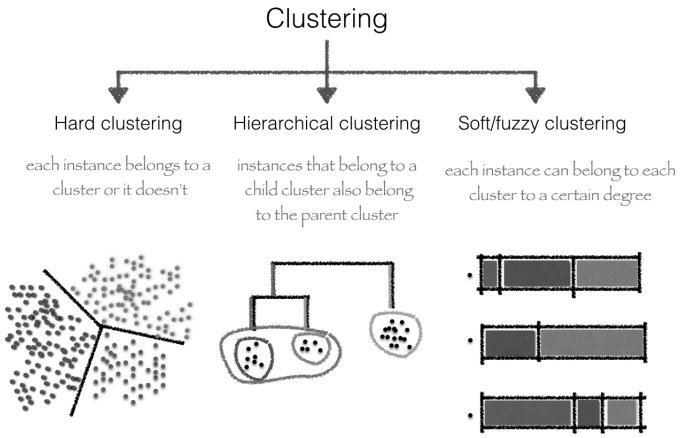



**Figure 2.2** An overview of types of clustering. From left to right: hard clustering, hierarchical clustering, and soft/fuzzy clustering. Adapted from the PyTexas conference 2015 in Texas, September 25, 2015. Retrieved September 4, 2018, from chdoig.github.io/pytexas2015-topic-modeling/#/2/5. Reprinted with permission. Image is licensed under CC BY 2.0 (Doig, 2015).

## 2.2 Clustering

In their review about data mining techniques for driving data, Constantinescu, Marinoiu, and Vladoiu (2010) evaluated hierarchical clustering analysis (HCA). The authors define HCA as: "HCA classifies the drivers according to some variables so that homogeneity within and heterogeneity among groups are obtained" (Constantinescu et al., 2010, p. 658). More specifically, hierarchical clustering uses a proximity measure to cluster data points which contain the lowest distances. Subsequently, this process is repeated, until one cluster of the complete dataset is created. The middle visual representation in Figure 2.2 is an example of an abstract hierarchical cluster analysis, in which each branch is represented by one cluster. By combining smaller clusters, larger clusters are created. The main advantage of HCA was mentioned in the study of Farrelly et al. (2017), in which HCA "minimizes the introduction of error and better preserves the local and global structure of the data" (Farrelly et al., 2017, p. 95). This allows researchers, who explore naturalistic driving data, to research broader constructs in data. For example, clusters high in hierarchy, could potentially detect aggressive driving behavior, while clusters lower in hierarchy, might be able to highlight distinctions between aggressive drivers. Another advantage of HCA, which Farrelly et al. (2017) revealed, is the ability to analyze naturalistic driving data without sample size limitations. However, in a review about HCA, Kumar, Dhok, Tripathi, and Tiwari (2014) mentioned that in order to receive stable results from HCA, noisy data and outliers need to be excluded from datasets, as HCA might perform worse including these. Therefore, HCA

might not be the optimal algorithm in analyzing naturalistic driving data, as inclusion of subtile differences (i.e., noisy data and outliers) is key in analyzing driving behavior.

The third cluster algorithm which is discussed is soft clustering, also referred as fuzzy clustering. Soft clustering analyses samples in datasets, in which each sample is shaped in one or more clusters. Subsequently, each cluster has a specific degree of importance in each sample. Results in soft clustering are received by finding latent connections between samples in the dataset. The visual representation on the right in figure 2.2, provides an abstract overview of three samples in a dataset, which consist of three different types of clusters at most. The strength of the clusters is visualized by the length of each bar in each sample. Bora, Gupta, and Kumar (2014) mentioned in a comparative study between hard and soft clustering, that flexibility is the main advantage of soft clustering, as samples can be divided in more than one cluster. Another advantage of soft clustering is the ability to find underlying information in shape of latent variables. More specifically, latent variables analyze the collection of samples, and return the underlying meaning of the data. One unsupervised algorithm, which includes latent variables, is Probabilistic Topic Modeling (PTM) (Blei, 2012). Initially, PTM was designed for text mining analyses, but recently it has been applied for other data mining purposes, such as e-Health data, financial data, and naturalistic driving data (J. H. Chen, Goldstein, Asch, Mackey, & Altman, 2017; Venkatraman, Liang, McLaurin, Horrey, & Lesch, 2017). Both studies share in common that they used their field-specific data analogously to text documents. Venkatraman et al. (2017) examined data of patients who were treated within the first 24 hours, and included textual data about their initial information and the doctors prescription. By applying PTM they were able to create a topic model, which was able to automatically create topics which would benefit decision support content, thus, allowing health care to function more efficiently. Venkatraman et al. (2017) measured differences in expected and unexpected events involving crosswinds. The researchers applied PTM to include variables such as, steering angle, and the frequency of changes in steering angle. Besides that, the researchers were able to translate time-series data to symbolic representation. The next section will describe the steps which are required to transfer time-series data to symbolic representation.

## 2.3 Symbolic Aggregate Approximation

In their study, McLaurin et al. (2014) mentioned that analysis of time-series data is challenged, as large amounts of data make it difficult to maintain the overview of the data. McLaurin et al. (2014) refer to this as "compre-

hensive analysis", and emphasize that data reduction or data compression is required to maintain the synopsis of large amounts of data (p. 2107). To complement comprehensive analysis, studies experimented with high-level representations, such as "Fourier transforms, wavelets, eigenwaves, piecewise polynomial models" (Lin, Keogh, Wei, & Lonardi, 2007, p. 107). However, Lin et al. (2007) argued these methods did not treat dimension reduction of data, which makes analysis with data mining techniques problematic. Moreover, previous methods applied distance metrics to convert data to symbolic representations. However, this type of conversion changed the original data structure, and was therefore, not adequate enough to apply for future analyses (Lin et al., 2007). To cope with this problem, Lin et al. (2007) introduced Symbolic Aggregate Approximation (SAX). SAX is capable to turn time series data into symbolic representation, while containing the data structure. The conversion of time series data consists of two steps. First, each instance in time series data is transformed to a alphabetical representation. This forms part of Piecewise Aggregate Approximation (PAA), in which ranges in time series data are divided into equal sized window frames. This is achieved by normalizing the data, and to divide time frames by the probability of occurrence of each window frame (see figure 2.3). Each instance in the time series data is then assigned to in one of the alphabetical representations, which the PAA conversion has provided. The second step in PAA conversion is to combine the alphabetical transformation of words of one time series occurrence into a string.

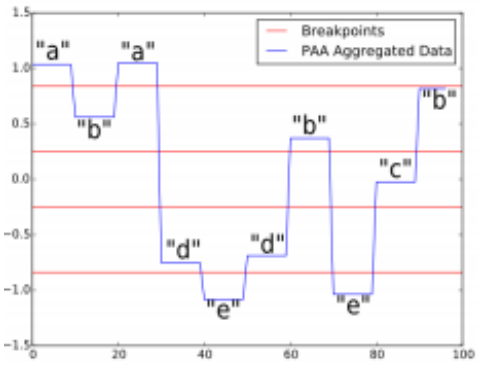



**Figure 2.3** Conversion of time series to SAX-representations. Time series are first normalized. Each frame in this figure has an equal size, thus, representing the probability of each symbol occurring. Each frame is divided by breakpoints. Then, values in the time series are converted to the alphabetical representations depending in which frame they coincide. The numerical conversion, in this case, is translated as "abadedbecb" (Puschmann et al., 2018, p. 5).

Lin et al. (2007) were able to apply SAX-conversion on time series, and

created a methodology to reduce the dimensionality of data. Their method, enabled consecutive studies not only to analyze large amounts of time series data, but also to combine various variables (McLaurin et al., 2014; Puschmann et al., 2018; Venkatraman et al., 2017). For example, Puschmann et al. (2018) applied traffic data, and combined this with weather data, which was available in the city of Aarhus in Denmark. Venkatraman et al. (2017) were able to analyze expected events during driving behavior of participants. Steering wheel angels and steering rate were combined using the SAX-methodology. McLaurin et al. (2014) combined acceleration, and speed data, to construct a SAX-representation, which was used for Probabilistic Topic Modeling, which will be discussed in next section. In conclusion, SAX-methodology has enabled a variety of research fields the opportunity to include large amounts of time series data, which can also be combined. At the same time, time series data contains its structure, but dimensionality is reduced, which allows SAX-represented data to be analyzed with data mining techniques.

## 2.4 Probabilistic Topic Modeling

In his short review about PTM, Blei (2012) mentioned the usability of PTM in textual analysis. A major advantage of PTM, which was mentioned, is the convenience of PTM-algorithms to automatically structure datasets (Blei, 2012). In a text mining perception, PTM is able to analyze collections of documents and to uncover all hidden topics. Besides that, PTM finds the structure of topics in each document, by indicating the presence and location of topics in documents. LDA is described as "generative probabilistic model of a corpus" (Blei, Ng, & Jordan, 2003, p 996). More specifically, LDA analyses documents in a corpus, and assumes that these documents are a mixture of topics, with topics consisting of collections of words. Each topic is then a collection of words, which indicate their strength in a topic by means of a probability. To conduct LDA, three assumptions need to be included. First, LDA expects a determined $N$ of words in a document, based on Poisson distribution. Second, a topic mixture must be chosen by determining $\theta$. Third, each word in the document must be generated by (1) choosing a topic ($z$), which is assumed to exist in a corpus by a multinominal distribution, and (2) by choosing the topic that generates words ($w$), based on the multinominal distribution of the topic (Blei et al., 2003). Subsequently, the generative process of LDA is expressed in the following equation:

$$p(\beta,\, \theta,\, z,\, w_D) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \Big( \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p w_{d,n}|\beta,\, z_{d,n}) \Big)$$

LDA expects to receive at what topics the corpus consist of. With this information LDA returns the topic distribution per document, and word occurrences in topics. Blei et al. (2003) refer here to computing a posterior distribution of hidden variables. The equation, which applies to this is as follows:

$$p(\beta, \theta, z|w_D) = \frac{p(\beta, \theta, z, w_D)}{p(w_D)}$$

"LDA makes some important assumptions about corpora" (Remmits, 2017, p 5). LDA assumes documents are a bag-of-words representation, and thus, syntax are not taken into account. LDA works backwards when analyzing documents to identify the topics, and word distributions in topics. LDA achieves this by learning from topic representations in text. Words in documents are first randomly assigned to one of the predefined topics. Then, LDA calculates the proportions of all words, which are present at that moment in that topic. Subsequently, the proportion of that topic over all documents is calculated. After iterating until words in the corpus are assigned to topics, and topics to documents, the LDA-model is created.

LDA has been proven to be an effective tool in studies as it is an easy to understand algorithm, while creating precise results (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). Furthermore, LDA is able to detect semantic coherent topics without human intrusion. Therefore, LDA has improved text mining related studies to quickly and effectively determine topics in text. However, a downside of LDA is required input of $K$ topics (Blei, 2012). Thus, in case of studies, which explore topics in large corpus, a trial-and-error approach of topic determination is expected to find to optimal set of parameters for the eventual LDA-model.

## 2.5 Psychomotor Vigilance Test

The current research is interested in differences in driving behavior between alert and unalert individuals. McLaurin et al. (2014) differentiated between OSA-patients and non-OSA-patients. An effect of OSA-patients on a daily life is the presence of daily fatigue. Subsequently, fatigue has a negative influence on the alertness-state of an individual (Song et al., 2017). In the study of Song et al. (2017), the researchers tested alertness on different type of participants, who were categorized by age. Younger participants were significantly affected by fatigue and would therefore, perform worse on a alertness-maintaining task. Thus, detection of fatigue among drivers is essential in order to determine whether their personal state influences their alertness (Bener, Yildirim, Özkan, & Lajunen, 2017).

To measure the alertness state of participants in a driving simulation, multiple studies have included the *Psychomotor Vigilance Test* (PVT) (Aryal et al., 2017; Brunnauer et al., 2018). PVT is a reaction time test in which individuals are randomly exposed to stimuli over an period of time ranging from 3 to 10 minutes per test. During a test, an inter-stimulus appears on a screen which irregularly is displayed between two and ten seconds, and participants are requested to react as fast as possible to the stimulus.

PVT has allowed studies to analyze the effect of sleep loss on individuals Loh, Lamond, Dorrian, Roach, and Dawson (2004). The objective of studies which implemented PVT, is to measure whether a decrease in reaction time can be measured, and thus, depreciation of psycho-motor skills. Studies have successfully found differences in reaction times as errors increased and vigilance decreased as a result of sleep deprivation.

Originally, PVT required to include ten minute intervals per test to ensure validity of alertness results. As researchers argued intervals could be diminished while ensuring similar results, lower intervals of three to five minutes intervals were tested. Researchers were able to confirm five minute intervals in PVT, which led to similar results as to ten minute intervals (Jones et al., 2018). A major advantage of lower intervals is the practicality of conducting experiments with PVT in situations were surroundings are hectic. Besides improvements in time intervals, studies have also measured the effect of conducting PVT on different devices (e.g., mobile phones, tablets). In the past, PVT required to be conducted on a computer. Arsintescu et al. (2017) studied the difference between computer-based PVT and touchscreen devices, and concluded that the results for both tests had no significant differences. Therefore, their study contributed to the validity of hand-held devices with touchscreen, and allowed future studies to apply PVT in more hectic surroundings.

PVT allows to measure the alertness state of individuals. In the experimental setup of McLaurin et al. (2014), participants were selected for research due to their health-state (i.e., OSA patients versus non-OSA patients). An attempt in this study is made to replicate this study (McLaurin et al., 2014). However, due to resources the current study did not have access to a human subject pool with OSA-patients. On the other hand, Batool-Anwar et al. (2014) were able to apply PVT in their experiments and were able to confirm that PVT was useful for assessments of fatigue along individuals. Moreover, PVT allows more reliable results than subjective sleep evaluations (Batool-Anwar et al., 2014). To distinguish between alertness states, thresholds were determined to classify participants in alert versus unalert states. Chapter 5 describes how PVT was conducted during experiments and how thresholds were determined.

# 3 Experimental Setup Driving Experiment

This section establishes the procedures which were performed to create the desired Probabilistic Topic Model (PTM). First, this section describes the collection of the data. After this, the preprocessing steps are described and the experimental setup to conduct PTM is further explained.

## 3.1 Data collection

### 3.1.1 Participants

In total, twenty-five participants were recruited through convenience sampling (5 women, 20 men, $M_{age} = 28.38$, $SD_{age} = 8.42$, age range: 19–59)[1] As participants were closely available to the researcher, they were willing to participant on a voluntarily basis.

### 3.1.2 Design

The study applied an experimental design, where all participants would perform a similar experiment. That is, each participant was exposed to one condition.

### 3.1.3 Apparatus and materials

For this research, a telematics sensor, type T31 was applied (see figure 3.1). This device, which was provided by Crossyn, was installed on the battery of the vehicle. The T31 was powered by the car battery. The sensor consisted of a GNSS receiver, which was connected to a GNSS satellite. The connection between receiver and satellite, made it possible to determine the position and timestamps of the vehicle. Subsequently, datapoints made it possible to determine the velocity of the vehicle.

Participants completed their task in a 2017 Toyota CH-R vehicle with manual transmission (see figure 3.2). This gasoline powered vehicle was exclusively used for all tests in this study. Toyota Netherlands granted the use of this vehicle between December 4th and December 11th, 2017.

[1] $M$ = mean, $SD$ = standard deviation.

**Figure 3.1** T31-sensor

**Figure 3.2** Toyota CH-R

### 3.1.4 Procedure

Twenty-five experiments were conducted over a period of six days from December 4 until December 9, 2017. During registration, participants indicated time slots and pick-up point, which depended on convenience in daily schedule of the participant. Participants were asked to refrain from being intoxicated at least 12 hours before the experiment as this could have influence on the experiment. To ensure safety, experiments did not take place if one or more health situations applied to a participant. All participants were generally in good health and mental condition. Participants received a consent form and were asked to give formal consent in case they agreed with the conditions of the experiment. After formal consent was provided, participants drove one session of approximately 30 minutes. A timer started counting as soon as the participant and researcher drove away from the starting point. The timer was stopped at least after 30 minutes and stopped as soon as the driver reached the end point and switched of the engine.

### 3.1.5 Driving task and environment

Besides formal consent, participants received oral instructions for clarity of the driving task. Participants were instructed that they would receive oral directions throughout the driving task, meaning they did not have to nav-

igate themselves. Visual navigation was excluded from the experiment, as drivers were required to stay focused on the road to ensure safety. Also, auditory input in the vehicle was diminished by keeping the radio mute [2]. Finally, participants were allowed to have a casual conversation during the experiment in order to imitate natural driving behavior.

After instructing each participant, the experimenter, who took place next to the driver's seat, started to give directions. The routes in the experiments were not identical as start and end points differed based on what participants indicated in their registration. Although the routes were mostly not identical, most of the routes showed similarities, as the experimenter maintained the city ring in the city of Tilburg in the Netherlands, which covered approximately 30 minutes. Retaining this route ensured the driving task of sufficient driving time. As figure 3.3 illustrates, the city ring in Tilburg is represented by long uneventful roads, where the speed limits vary from, 50 km/h to 70 km/h.

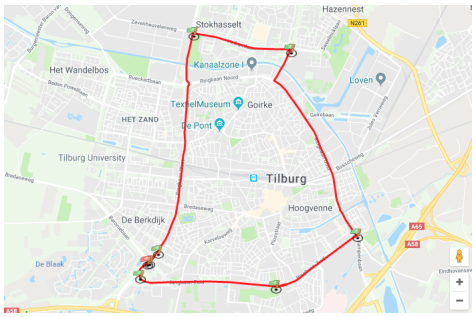



**Figure 3.3** City ring in Tilburg, The Netherlands. The red line indicates the predefined route held during the experiments. Detailed routes are excluded from this figure to protect anonymity of participants.

## 3.2 Dataset

### 3.2.1 Data overview

The dataset, which was provided by Crossyn, consists of 24 rides, which were recorded from December 4th until December 9th, 2017. One ride was excluded from the dataset, as the T31 lost its connection due to an unknown reason. Before delivering the data, a Crossyn software engineer interpolated the data between instances, as the T31 illustrated infrequent gabs in seconds

[2] In their study, Brodsky, Olivieri, and Chekaluk (2018) concluded that hostile music can lead to distracted drivers.

between instances. More specifically, interpolation made it possible to fill in missing values. This was done by interpolating data within instances that were two or more seconds apart from each other. In total, the experiment stored 813.30 minutes of driving data, which equals 13.55 hours. Every experimental task took approximately 30 minutes per ride (MPR, $M_{ride}$ = 33.89 MPR, $SD_{ride}$ = 8.33 MPR).

### 3.2.2 Preprocessing and Feature Extraction

All preprocessing and analysis steps were applied in Python 2.7 using PyCharm (Professional 2018.1 version) as a software tool. First, driving data was set to SAX by utilizing the Python package **saxpy.alphabet**[3]. Similar to the study of McLaurin et al. (2014), at most 8 alphabetical representations of speed and acceleration were defined. Table 3.1 illustrates the alphabet used for this study, and its corresponding ranges.

| | Range | | | |
|---|---|---|---|---|
| SAX letters | Speed (km/h) | *N* | Acceleration (g) | *N* |
| a | 0.0 to 1.0 | 10566 | −1.1890 to 0.0760 | 3728 |
| b | 2.0 to 16.0 | 4161 | −0.0708 to −0.0472 | 2474 |
| c | 17.0 to 28.0 | 4407 | −0.0437 to −0.0283 | 6224 |
| d | 28.7 to 38.0 | 5417 | −0.0212 to −0.0094 | 56 |
| e | 39.0 to 48.1 | 10513 | −0.0081 to 0.0071 | 24230 |
| f | 48.4 to 59.3 | 4011 | 0.0089 to 0.0239 | 62 |
| g | 60.0 to 74.0 | 3054 | 0.0283 to 0.0438 | 5208 |
| h | 75.0 to 124.0 | 6669 | 0.0453 to 0.0708 | 3000 |
| i | n.a. | n.a. | 0.0755 to 1.2180 | 3815 |

**Table 3.1** Conversion of continuous input to SAX output.

After defining the alphabet, letters were combined into words. The structure of each word consisted of one letter (i.e. from "a" to "h") which coincided with speed. Subsequently, the second letter in the converted words, coincided an acceleration-letter from "a" to "i". Lastly, the letters were joined by placing an underscore between letters of each word. In total, 67 unique words were created.

The current study exhibits similarities with the frequency of speed letters in the study of McLaurin et al. (2014), partially reproducing their results, but in another context, since this study aims at identifying distinctive behavioral profiles in a general setting. For readability purposes, letters which present speed intervals are mentioned by a capital letter and acceleration intervals by lower case letters. Table 3.1 indicated the most occurring letters in the data set were "A" ($N = 10556$), "E" ($N = 10513$), and "H" ($N = 6669$). As

[3]https://github.com/seninp/saxpy

"A" stands for a very low speed, it represents the car being in a stationary position. The letter "E" in its turn is represented by a speed between 39.0 and 48.1 km/h, which results in a modest speed. This speed coincides with maximum speed barriers between 30 to 50 km/h, and can be defined as city driving McLaurin et al. (2014). Followed by the "E", the letters "F", and "G" range between speeds of 48.4 to 74.0 km/h, which present roads where speed limit is 70 km/h. Hence, this driving behavior can be defined as moderate. The letter "H" ranges from 75.0 to 124.0 km/h. This speed range is similar to the study of McLaurin et al. (2014), in which the letter "H" was defined as high speed driving. Therefore, the current pre-processing steps indicate similarities to this previous study.

## 3.3 PTM

The SAX corpus which was constructed in the previous section formed the vocabulary for PTM. Then, the topic modeling library **scikit-learn**[4] was used to prepare the analysis. Subsequently, LDA required a document word input. By applying *Countvectorizer* the representation of SAX words in the text was simplified by an object, which stored the occurring words in the following format: *(0, 1) 2*. The first number in the tuple presents the number of the document, which is the first document in our dataset. Before assigning a number to all words in the corpus, countvectorization ordered all words in alphabetical order. Hence, the first word that occurred in the dataset was represented by the word *"A_a"*, and its number was assigned as *1*. The third number outside the tuple was the frequency of the word in the document. Thus, *"A_a"* occurred two times in the first document. The vectorization of the dataset was further applied on the corpus (i.e., documents and words). The collection of words in documents results in a *bag-of-words* (BOW), which is applied in natural language processing to visually represent words by means of vector space.

The next step after constructing the BOW, was to apply it on LDA. In order to find the topic model with the best fitting topics, a grid-search was applied. The parameters which were applied were: number of topics and learning decay. LDA requests for input for the amount of topics. Therefore, a range from 1 to 5 (inc. 5) was given as input. LDA, which is created using the scikit-learn library, consists of Online Variational Bayes algorithm, which is applied in large amounts of documents (Hoffman, Blei, & Bach, n.d.). Learning decay is applied to control the learning rate. The values, which were set for this range, vary from 0.5 to 1.0. For this grid-search, three inputs were applied (i.e., 0.5, 0.7, 0.9). The default for learning decay is set at 0.7. After parameters were defined, the LDA-model was

[4]https://github.com/scikit-learn/scikit-learn

generated by applying a grid-search class in scikit-learn. The input for this model consisted of determining which model had to be constructed (i.e., LatentDirichletAllocation()) and with which parameters. The model was fitted by applying the vectorized data.

## 3.4 Parameter optimization

The current study was interested in evaluating potential differences in behavioral driving topics. Unlike the study of McLaurin et al. (2014), in which no strategies were applied to improve the LDA-model, this study will optimize the LDA-model, by including parameters that analyze the corpus of this study in various ways. The following text-mining strategies were included:

- ***n*-grams:** were applied in the Countvectorizer ranging from 2 to 3 words. Hence, for the analysis bi-, and trigrams were included.
- **max_features:** is a function in NLP, which allows researchers to indicate the maximum amount of features (i.e., words) to include in the BOW. For this study, *max_features* was set at 100, meaning LDA did include the top 100 most used words in the corpus, and excluding all other words.
- **min_df and max_df:** can be set to include and exclude words which occur too frequently in text. The default for min_df is 1, meaning that words are excluded which appear only in one document. The default for max_df is 1.0, which means that no words are ignored in the full corpus. In this study, a max_df of 0.70 was selected, which means that words were excluded which appeared in more than 70% of the documents. Applying these parameters allows to include words in the corpus, which did not occur in most documents, and therefore, potentially finding subtile changes.

# 4 Results Driving Experiment

This section presents the results of the LDA experiment, which was created using naturalistic driving data from twenty-four rides. The LDA model was created by empirically analyzing the parameters, which determined the optimal topic model.

## 4.1 Experimental results

As a first experiment, an LDA model was created by leaving out text mining strategies. Section 2.4 mentioned that LDA expects to receive $K$ topics in order to create a model around this. In this study, $K$ topics were not provided as human observation in driving data is challenged. The goal to find the optimal number of topics, is not to include topics with redundant information. In order to determine $K$ topics, learning decay was assessed. Learning decay indicates which point is optimal in a sense of learning rate. Preferably, a learning decay close to 0.0 is desired. Figure 4.1 illustrates the results of the learning decays. In order to choose the best fitting model, a grid search approach was included. The results illustrate that this maximization is achieved with the learning decay of 0.5, meaning it outperforms both 0.7 and 0.9, with a learning decay of approximately -50,600. Hence, for this topic model, three topics were created.

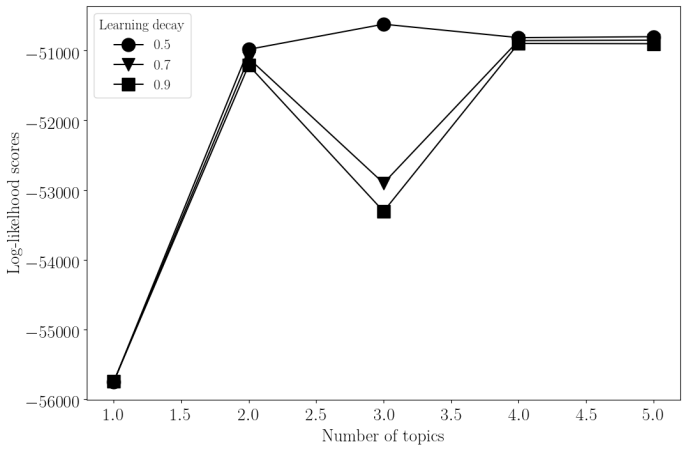



**Figure 4.1** Choosing optimal LDA model by assessing the learning decay. The optimal LDA-model is created with three topics.

## 4.2 LDA model

The experiment has resulted in a fitted topic model, which consists of three topics. This section will discuss the topic definition, the topic distribution along the dataset, and a visual representation using the Python package pyLDAvis[1].

### 4.2.1 Topic definition

The LDA model has created a topic model of three topics in total. Each topic consists of the probability of keywords which explain the weight of significance. The top keywords are extracted from the topic model by converting the vectorized dataset to the featured names. This leads to an overview of SAX words, which describe the composition of each topic. Table 4.1 illustrates the output of each topic by showing the words with the highest weight. Each topic will be further described in the next sections.

| | Topics | | |
|---|---|---|---|
| Words | Topic 1 | Topic 2 | Topic 3 |
| Word 1 | A_e | A_e | H_e |
| Word 2 | E_e | E_e | A_e |
| Word 3 | H_e | E_c | H_c |
| Word 4 | E_g | E_g | H_g |
| Word 5 | G_e | F_e | E_e |

**Table 4.1** List per topic of top five SAX words with highest weights in descending order.

#### 4.2.1.1 Topic 1: city driving

Topic 1 has the strongest weight for word "A_e", which is characterized by the letter "A" and "e". Previously, table 3.1 described the ranges in which each letter coincided, meaning that "A_e" explains the car in a stationary position. The second strongest word (i.e., "E_e") identifies city driving with very low acceleration. Third, the letter "H_e" defines a higher constant speed during a ride (i.e., 75.0 and 124.0 km/h) with very low acceleration. The fourth word in topic 1 is "E_g", which describes a city driving speed with a higher acceleration. Last, the word "g_e" describes a high constant speed (i.e., 60.0 to 74.0 km/h) and again with very low acceleration. Except for word 4, the top occurring words describe a constant speed between 39.0 and 48.1 km/h. Inclusion of the word "H_e" with a speed varying between

[1]https://github.com/bmabey/pyLDAvis/blob/master/pyLDAvis/sklearn.py

75.0 to 124.0 km/h deviates from other words in this topic. However, it might be that the original data in this topic consists of speeds closer to 75.0 km/h than 124.0 km/h. Furthermore, "A_e" is a frequent occurring word in the dataset, and is listed on top of the other words. Thus, topic 1 is described as city driving.

#### 4.2.1.2 Topic 2: complete city driving

Topic 2 has similarities when compared to topic 1. This indicates that topic 2 is similarly clustered as city driving. The difference in topic 2 are the words "E_c" and "F_e". First, "E_c" can be described as a speed between 39.0 to 48.1 km/h with deceleration, meaning that speed of the car represents city driving, and is decelerating at the same time. This action will most likely occur right before the car is about to turn into stationary position. The second word, "F_e", which also deviates from topic 2 is represented by a constant speed between 48.4 and 59.3 km/h and a very low acceleration. As topic 2 differs by the two most occurring words, it resembles city driving similarly to topic 1. However, the top 5 words of topic 2 describe city driving more thoroughly as it includes a speed between 48.4 and 59.3 km/h. The maximum speed of roads which were included during the experiment were mostly 50 km/h. This word is essential in the definition of city driving. Therefore, topic 2 subsists of a more complete overview of city driving.

#### 4.2.1.3 Topic 3: Highway driving

Topic 3 clearly highlights different top words compared to topic 1 and 2. The highest ranked word, "H_e", represents a constant speed between 75.0 to 124.0 km/h, and with very low acceleration. Then, topic 3 includes word "A_e", which is a stationary position of the car. Third, "H_c" represents a similar speed range as word 1, but instead it denotes a decelerating state. Fourth, "H_g" indicates the same speed range as the word "H_c", but rather than deceleration, this word is a representation of acceleration. Lastly, the word "E_e" is a representation of a constant speed range from 39.0 to 48.1 km/h. As the top words give an indication of higher speeds, this topic can be defined as highway driving.

### 4.2.2 Topic representation in documents

An essential part of LDA analysis is to distinguish which topics belong to which documents. Table 4.2 provides an overview of all documents which were included in this study. Then, weights of topics indicate to which extend a topic is represented in each document. The weights of all topics, which are included in one document, accumulate to 1.0.

In table 4.2 Topic 2 and 3 are included, and Topic 1 excluded. The reason for this is that topic 1 had no occurrence in any of the rides. Interestingly, LDA analysis found the best fitted model with three clusters, but topic 1 shows no significance when assigning topics to rides. Section 4.2.1 highlighted two topics, which in essence appeared to be similar. The results in Table 4.2 confirm the redundancy of topic 1 compared to topic 2 and explain why topic 1 has no significance in the distribution of topics.

| | Topics | |
|---|---|---|
| Document number | Topic 2 | Topic 3 |
| 1 | 0.98 | 0.02 |
| 2 | 1.00 | 0.00 |
| 3 | 1.00 | 0.00 |
| 4 | 1.00 | 0.00 |
| 5 | 0.34 | 0.66 |
| 6 | 1.00 | 0.00 |
| 7 | 1.00 | 0.00 |
| 8 | 1.00 | 0.00 |
| 9 | 1.00 | 0.00 |
| 10 | 1.00 | 0.00 |
| 11 | 0.86 | 0.14 |
| 12 | 0.94 | 0.06 |
| 13 | 0.00 | 1.00 |
| 14 | 0.00 | 1.00 |
| 15 | 1.00 | 0.00 |
| 16 | 0.00 | 1.00 |
| 17 | 1.00 | 0.00 |
| 18 | 1.00 | 0.00 |
| 19 | 1.00 | 0.00 |
| 20 | 1.00 | 0.00 |
| 21 | 0.14 | 0.86 |
| 22 | 1.00 | 0.00 |
| 23 | 1.00 | 0.00 |
| 24 | 1.00 | 0.00 |

**Table 4.2** Weights of topic occurrences per ride for topic 2 and 3. Topic 1 is excluded from the overview as no occurrences were present for this topic in the experimental sample.

In order to give insights which topic was most dominant in each document, a topic distribution is displayed in table 4.3. A surprising result is that 5 rides are clustered as highway driving, which correspond to the rides that were held in the driving experiment. Thus, LDA analysis succeeded in correctly clustering the documents into the type of driving which was most dominant during each ride.

| Topic number | Number of documents |
|---|---|
| 2 | 19 |
| 3 | 5 |

**Table 4.3** Distribution of dominant topics in current dataset.

### 4.2.3 Visualization topics

The aforementioned python package pyLDAvis, was applied to the LDA analysis. This function in Python allows to visually and interactively represent the topics in HTML. Figure 4.2 illustrates the interactive output from the LDA model. On the left, the bubbles represent the topics in a semantic topic space. This means that the closer these bubbles are to each other, the more semantic resemblance they share. Figure 4.2 indicates that topic 2 and 3 do not share common words, as they appear on a long distance from each other on the distance map.

On the right side of Figure 4.2 the words are displayed which were applied to the LDA analysis. The interactive visualization makes it possible to highlight a word. Subsequently, the sizes of the bubbles on the left adapt to the prevalence of the word inside the topic. This means that the more important a word appears to be in a bubble, the larger the size of the bubble.

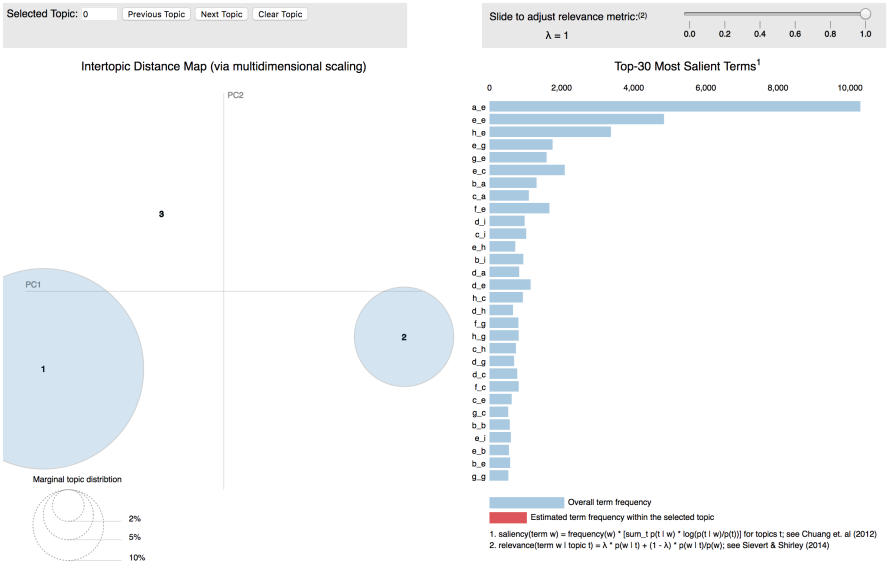



**Figure 4.2** The topics from the LDA analysis are visualized by applying pyLDAvis. Topics are represented by bubbles on the left side indicating their sizes and respective distances to each other as obtained by multi-dimensional scaling. The right side of the figure displays the word term space, visualizing the respective term frequencies. The words are ranked in descending order of importance.

## 4.3 Different strategies in Topic Modeling

In the following, steps are outlined for identifying behavioral driving patterns using a more complex topic modeling approach. As previously discussed in Section 3.3, these refinements include using (1) $n$-grams (bigrams, trigrams), (2) a restricted set of $n$-grams, and (3) a selection of $n$-grams forming the topic models.

### 4.3.1 Behavioral Topic Modeling using $n$-grams:

The first experiment applied bi- and trigrams on the data set to enable more complex patterns of words, and thus to potentially find stronger patterns than only applying unigrams. The most optimal learning decay was achieved with four topics. An overview of each topic, with most occurring vocabulary is shown is Table 4.4. In order to investigate if the results of this LDA-model differ from the LDA-model, which was described in Section 4.2.1, we assess each topic individually:

- **Topic 1 (highway driving)** distinguishes itself by high speeds. Section 4.2.1 defined highway driving as a topic, as speeds were included

which were higher than 75.0 km/h. Applying bi- and trigrams indicates that a combination of "H_e" determines the most important word in topic 1. More specifically, the highest proportion of "H_e" resides this topic. Thus, this LDA-model has defined topic 1 as highway driving.

- **Topic 2 (high way driving during rush hour)** is the smallest topic in this LDA-model. The speed in topic 2 ranges from minimal (i.e., "A") to a higher speed ("H"). Overall, term frequency is very low in topic 2, indicating the size of this topic is very small. More interestingly, topic 2 consists of a combination of high speed, and a stationary position. Referring to the driving experiments held for this study, one participant was subject to rush hour, while driving on a high way. As one participant was subject to this occasion, the size of topic 2 is explained.

- **Topic 3 (city driving with subtile changes)** indicates speeds which do not exceed the letter "H" (> 75.0 km/h). An interesting observation is the bigram "E_e D_c", meaning that the car shifted from a constant speed between 39.0 and 48.1 km/h to 28.7 and 38.0 km/h. Moreover, the acceleration decelerates to -0.04 and -0.03. A subsequent event which occurs in topic two is shown with trigram "E_e D_c D_e", which clearly indicates a change in constant speed from approximately 50 km/h to 30 km/h. Thus, topic 3 shows a subtile change in speed.

- **Topic 4 (city driving)** forms the largest topic in the LDA-model. The most relevant words indicate similarities with the previously established LDA-model in Section 4.2.1. More specifically, the top bigrams, which are shown in Figure 4.4, are defined as "A_e A_e" and "A_e A_e A_e". As previously explained, "A_e" is a representation of a stationary state of the vehicle. The third and fourth most important words of topic 4 are again a homogeneous combination of bi- and trigram, but instead of the word "E_e". The symbolic representation of speed and acceleration was previously described in Table 3.1. "E_e" was an indication of a constant moderate speed (39.0 to 48.1 km/h). Compared to Section 4.2.1 topic 4 can be determined as city driving as speeds do not exceed speed letter "H", which is higher than 75.0 km/h.

| | Topic Vocabulary | | | |
|---|---|---|---|---|
| Rank | Topic 1 | Topic 2 | Topic 3 | Topic 4 |
| 1 | H_e H_e | H_e H_e | A_e A_e A_e | A_e A_e |
| 2 | H_e H_e H_e | H_e H_e H_e | A_e A_e | A_e A_e A_e |
| 3 | H_c H_c | A_e A_e A_e | E_e E_e | E_e E_e |
| 4 | H_g H_g | A_e A_e | F_e F_e | E_e E_e |
| 5 | A_e A_e | F_e F_e | E_e E_d | F_e F_e |
| 6 | A_e A_e A_e | E_e E_e | F_f F_e | E_c E_c |
| 7 | H_e H_c | F_e F_e F_e | F_e E_c | G_e G_e |
| 8 | H_e H_c H_c | E_e E_e E_e | E_e D_c | E_g E_g |
| 9 | H_c H_c H_c | G_e G_e | E_f E_c | G_e G_e G_e |
| 10 | H_e H_e H_c | E_c E_c | E_e E_c E_e | B_a B_a |

**Table 4.4** Presentation of most occurring words for each topic in LDA-model with inclusion of bi-, and trigrams.

As implementing $n$-grams in the LDA-model created more topics compared to the first experiment, which were surprisingly informative, $n$-grams were investigated in more detail. As a first step, $n$-gram were restricted to a set of the top 100 (max-features) in the analysis. This created an optimal LDA-model of 4 topics. As this "max-features model" only includes the most common features of a corpus, it is expected that this LDA-model will indicate more general patterns in the driving data.

- **Topic 1 (Constant speed driving):** In this LDA-model, topic 1 moderately represented, compared to other topics. The top 2 features, which are included (i.e., A_e A_e A_e, and A_e A_e), determine the over representation of a stationary position along the dataset. The next words, which are defined after these top 2 words, share similarities in a sense that speeds are constant (i.e., represented by the acceleration letter "e"). Besides that, speed letters *D, E, F, G,* and *H* indicate a high variety in speed ranges in topic 1. As the commonality in these words is represented by a constant speed, topic 1 can be defined as constant speed driving.

- **Topic 2 (City driving with max 75 km/h):** The largest representation of topics in all rides, is topic 2 as it occurs in 17 rides. Similarly to topic 1, the top 2 occurring words include the stationary representation (i.e, A_e). Subsequently, there is a variety in speeds and acceleration which define topic 2. Most importantly, speeds do not exceed 75 km/h, indicating topic 2 as city driving. Furthermore, acceleration letters vary from *a, c, e, g,* and *i*. This is a representation

of city driving, as drivers are constantly accelerating and decelerating due to vehicle density, and traffic lights.

- **Topic 3 (Highway driving):** Similar to the previous two experiments, in which highway driving was included, topic 3 is highly represented by high speeds, and thus, high way driving. In essence, speeds do not occur below speed letter *H*. At the same time, acceleration letters vary from acceleration to deceleration, in which the two most important words of topic 3 are represented by constant high speeds and low acceleration. Thus, as speeds above 75 km/h are over represented in varying accelerations, topic 3 is defined as highway driving.

- **Topic 4 (High way driving during rush hour):** Topic 4 is similar to topic 3, in a sense that higher speeds are combined with the stationary state of the car. In this topic, even higher speeds are recorded. The fact that word combinations with "A_e" exist, indicates a driving situation within a rush hour, as in these driving situations it is common to stand still on a highway due to a high amount of traffic. This result corresponds with one participant, who drove on a highway during rush hour in the afternoon.

### 4.3.2 Behavioral Topic Modeling using selected $n$-grams

The last experiment makes it possible to include and exclude features, which are under or over represented in the corpus. The optimal LDA-model in this setting consisted of two topics. Similar to the previous experiments, this experiment included bi- and trigrams. The default *min-df* is set at 1, meaning that no features in the corpus are ignored. *max-df* was set at 0.5, meaning that words which occur in 50% of the corpus are removed, to prevent the majority of words to be included. This approach allows to zoom into less frequent words in the corpus, and to focus on subtile changes in driving behavior.

- **Topic 1 (High and low acceleration and deceleration):** presents a wider variety of SAX-representations. For example, the first most relevant word in this topic is D_g D_g D_g, meaning an occurrence during a ride in which a driver would accelerate strongly, while finding itself between a speed of 28.7 to 38 km/h. Typically, in real driving situations, a state of D_g would occur in an situation where a driver would strongly accelerate to reach a point where the speed would be constant. In a city driving environment, a driver could accelerate to a point to which the speed would be *E* (i.e., between 39.0 and 48.1 km/h). The second, most important term in topic 1 is G_c G_c G_c. A close look at Table 3.1 reveals that this trigram explains a state in driving behavior in which the car is slowly deceleration, and in a

fairly high speed (i.e., between 60.0 and 74.0 km/h). The following terms in topic 2, are indications of high acceleration and deceleration. For example, the trigram "C_c C_c C_c" is an occurrence in which the vehicle is decelerating. Furthermore, D_a D_a D_a is a state in driving situations in which the vehicle is decelerated very strong, while the speed is between 28.7 and 38.0 km/h. All salient terms in topics of previous experiments, which included bi- and trigrams, and restriction using max-features, indicated mostly zero acceleration with SAX letter "e". The current topic sheds light to SAX-words, which represent a variation of acceleration and deceleration in driving behavior. Since 70% of most occurring bi- and trigrams in the corpus were removed, this time topics were shaped, which include less situations that occurred during the driving experiments.

- **Topic 2 (High speed driving):** the largest proportion of this topic consists of combinations, which include the SAX-word "H_e". Moreover, each combination in this topic is defined with the letter "H", which is the representation of speed between 75.0 to 124.0 km/h. Furthermore, the variation of accelerating symbolic representations is almost complete as the acceleration letters $b$ to $h$ are represented in the topic terms.

# 5 Experimental Setup Psychomotor Vigilance Test

To establish a framework which enables to determine the alertness state of participants, a Psychomotor Vigilance Test (PVT) was applied. The objective of PVT was to objectively quantify the vigilant state of participants prior and after the driving experiment, as it was hypothesized that driving a vehicle would have effect on individuals. The following section describe the procedure of PVT, which were conducted during the experiments, the data preprocessing, and determination of the alertness state of participants.

## 5.1 Procedure

During all experiments (i.e., prior and after the driving experiment) participants were exposed to PVT. In order to execute PVT, an Ipad Pro 12.9" (iOS 9.0) with a PVT application named Vigilance Buddies (available in the Appstore), was used for the experimental setup. Inter-stimuli appeared randomly between 2 and 10 seconds within a time range of 5 minutes. A visual stimulus, displayed in counting milliseconds, was shown after each inter-stimulus started. Milliseconds continued counting until a participant reacted, by means of simple reaction time, to the stimulus by tapping the touchscreen of the iPad. Simple reaction time is a requested response in experiments, where participants are exposed to one stimulus, and only one reaction is required (Sehgal & Kapoor, 2018). A participant was not required to tap a specific location on the iPad, but any touch on the screen would trigger the inter-stimulus to stop counting. After the trigger was recorded on the iPad, the point at which the milliseconds stopped accumulating, was displayed for 2 seconds. After this, the screen was reset to a black background until the next inter-stimulus started.

Participants who reacted too soon to a inter-stimulus, received immediate feedback, with a brief message (i.e., "False Start"). A *False Start*-indication was displayed in case participants responded before a inter-stimulus appeared up until <100 ms. Reaction times of <100 ms were previously determined as anticipated reactions of participants (Basner & Dinges, 2011). More specifically, Sehgal and Kapoor (2018) determined the average visual reaction time of individuals to be at approximately 209 ms, with a standard deviation of 42.50 ms. The results in the study of Sehgal and Kapoor (2018) did not require to adjust the modality of the current PVT. Reaction times of >500 ms were denoted as *True Errors*, which is a current determination

in PVT of errors that occur during attention inducement (Basner & Dinges, 2011). Before each experiment, participants were instructed to the PVT. Prior to every first test, a 1-minute pretest was conducted to familiarize participants with the process of the PVT.

## 5.2 Data Preprocessing of PVT

The results of each PVT were recorded in CSV-format and sent via e-mail to the experimenter. In total, $25 \times 2$ PVT were recorded. Two PVT of one participant were excluded from the dataset, as the data of the driving experiment were not valid for this participant. The reaction times (RT) in each PVT were divided in *False Start* (i.e., RT <100 ms, FS), *True Errors* (i.e., RT >500 ms, TE), and *Correct RT* (i.e., RT $\geq$ 100 ms and $\leq$ 500 ms, CR). Then, the RT of each 5-minute PVT were prepared to be analyzed. First, the amount of $n$ FS and $n$ TE were determined. The quantity of errors were distinguished as errors before, and after the driving experiment. Subsequently, errors were placed in a histogram to visualize the alteration before and after the driving experiment. Second, the remaining RT, denoted as Correct RT, were grouped per minute to analyze the alteration per minute before, and after the driving experiment. In order to visualize the results, the CR were placed in boxplots, which would make it possible to analyze significant differences.

## 5.3 Data annotation

In order to categorize the vigilant state of participants, a distinction of (1) Alert, and (2) Unalert participants was determined. The objective of annotating participants in one of two states was to measure if driving conditions had influence on the vigilant state of participants.

### 5.3.1 PVT-errors

As mentioned in Section 5.2 two types of errors were collected, and visualized in a boxplot to analyze the alteration before and after the driving experiment. Appendix A visualizes the quantity of errors which occurred during all PVT-experiments. In order to annotate whether a participant was affected after the driving experiment, the errors made were compared. An increase in errors (i.e., False Starts and True Errors) indicated an impaired vigilant state, and thus, an participant who had inclined alertness. On contrary, a decrease in errors, indicated an increase in vigilant state, which meant the alertness state of a participant improved after the driving experiment. This approach of labelling was based on the study of Aryal et

al. (2017) in which alertness of participants was based on the amount of errors construction workers made during a 2.5 hour experiment. Appendix B indicates the alertness denotation for all participants included in this analysis. In total, six participants were denoted as an unalert-state versus 18 alert participants.

### 5.3.2 PVT Reaction Time

After denoting the alertness level of participants by means of labels, the next step was to validate the results. In the previous section, participants were defined as Alert or Unalert, based on the quantity of errors. In this section, the valid PVT-results (i.e., correct reaction times), were analyzed to determine whether differences in performances occurred during the driving experiment. Appendix C displays, per participant and per minute, the reaction times by means of a boxplot. The objective is to determine whether error bars of boxplots, before and after the driving experiment, deviate by non-overlapping error bars. All results in Appendix C indicate that no significant differences were present as error bars overlap in case of all participants. These results will be further discussed in Section 7.

# 6 Results Topic Models versus Alertness-state

In this part of the study topic models, which were constructed in Section 4.3, are analyzed in combination with the labelled data from Section 5.2. In the study of McLaurin et al. (2014), the researchers were able to label participants as OSA or as Non-OSA patients. Subsequently, differences between these two groups could be analyzed in order to conclude whether driving behavior differed between these groups. The researchers based their results on differences in probability of topics between both groups. This section will present the results, based on the methodology of the study of McLaurin et al. (2014) and analyze if differences were found between Alert and Unalert participants.

## 6.1 Standard LDA-model

The first LDA-model, constructed in Section 4.2, was constructed with default settings and created three topics. In Appendix D all experiments are plotted with the objective to compare topic probabilities between alert and unalert participants. First, Appendix D.1a denotes a probability of 0.0, which conforms to the results previously illustrated in Figure 4.2 where no occurrences were found in Topic 1. The results in Topic 2 and Topic 3 are an indication of a similar variety in topic probabilities between alert and unalert participants. The difference between these two groups can be explained as a difference in the size of both groups. The size of the unalert population (i.e., 18 participants) was substantially larger than the alert population (i.e., 6 participants). This difference in population size explains the outliers, which can be found in Appendix D.1b and D.1c. The ranges in which the outliers of alert participants reside, compared to the error bars of unalert students are similar. Thus, no significant differences can be found between these groups.

## 6.2 Bi-and trigrams LDA-model

In this LDA-model, 4 topics were distinguished (i.e., (1) Highway driving, (2) Highway driving during rush hour, (3) City driving with subtile changes, and (4) City driving. The topics of this experiment will be further discussed in this section:

- In Topic 1, the distribution of highway driving is illustrated. As can be seen in Appendix D.2a, Topic 1 has a higher probability for alert participants compared to unalert participants. Interestingly, Appendix D.2a indicates 4 outliers in the alert group. This result resembles the experimental setup in which exactly 4 participants drove on highways. This might indicate that along the experiments, participants resided in an alert state. On the contrary, this result might also indicate that participants became more alert as participants drove on highways. At the same time, the results for unalert participants show that Topic 1 has low probability, meaning that unalert participants were not handling highway driving.

- Compared to the LDA-model of the first experimental setting in Section 6.1, Topic 2 did not occur in any of the experimental rides.

- Topic 3 indicated city driving with subtile changes. The experimental results indicate that Topic 3 only occurred in 2 rides, and only for alert participants. These results indicate that this Topic Model has created an extra topic, as two rides deviated from the majority of the corpus.

- In Section 4.3.1, it was mentioned that Topic 4 formed the largest topic in this Topic Model. The results in Appendix D.2d clearly illustrate that the weight of Topic 4 is the largest. However, three outliers, which represent alert participants, are shown in this figure. These outliers stand out in this figure, similarly as Appendix D.2a. Hence, the outliers which were present in Topic 1, show a resemblance in the fact that these rides were typisized by highway driving.

## 6.3 Max-features model

As the application of bi-, and trigrams resulted in interesting topics, these were included in the next two experimental setups to create LDA-models. First, a *max-features* model was created, which was set at max-features = 100. This setup created 4 new topics. The probability of these topics will be further discussed:

- Topic 1 was an indication of a large variety of speeds, which shared resemblance by the acceleration letter *e*, which combined with speed letters, signified constant speeds as acceleration was it its minimum. The results for Topic 1 in Appendix D.3a indicate a low probability of Topic 1 in both participants groups (i.e., Alert and Unalert participants). Two participants, labelled as alert, contained a high probability of Topic 1, meaning that during their ride, constants speeds were dominant.

- Topic 2, the largest topic in this LDA-model, represented city driving up tot a mamixum speed of 75 km/h. The results in Appendix D.3b illustrate the dominance of the probability of topics 2 for both types of participants. For both participants groups, the whiskers reach to a probability of 0.0, meaning rides in the corpus did not encounter driving behavior of Topic 2.

- Similar to the previous LDA-model in Section 6.2, the current LDA-model consisted highway driving. The results in this topic show few participants who were exposed to a highway setting. Appendix D.3c shows three clear outliers for alert participants, compared to 1 outlier for unalert participants.

- The latter topic, which was previously defined as high way driving during rush hour, is more represented by alert participants, compared to unalert participants. However, unalert participants had one outlier, with a high probability (i.e., $\pm$ 0.7).

## 6.4 Selected $n$-grams model

The last experimental setup consisted of omitting frequently occurring bi-, and trigrams (i.e., max-df = 0.7) and least occurring (i.e., min-df = 0.1). After applying this setup, 2 topics were extracted from the corpus.

- Appendix D.4a indicates the results for alert and unalert participants for varying acceleration situations. Alert participants are under represented in Topic 1 as the higher bound of the boxplot reaches less than a probability of 0.2. However, two outliers are found in this participant group. Unalert participants are, compared to alert participants, over represented in Topic 1. The topic probability of these groups ranges from 0.0 to 1.0, but with a mean closer to a probability of 0.0.

- Topic 2 in this setup was defined as high speed driving. The result in Appendix D.4b illustrates a reversed output compared to Topic 1. Here, alert participants have a high probability for Topic 1, and for unalert participants, the mean is closer to 1.0 than 0.0. For unalert participants this means that the probability of highway driving ranges from 0.0 to 1.0.

# 7 Discussion

The increase of road fatalities has led to the urge to find methods and systems to prevent risky driving behavior. Preferably, these methods need to be automated, as human interference might lead to a bias in risky driving detection. Furthermore, current techniques, which record naturalistic driving behavior, have brought the potential and challenge in detection of driving behavior. Large amounts of data being captured automatically have the potential to capture individual driving behavior, which inform us about how people drive. Therefore, the first question in this study sought to determine which topics can be distinguished from naturalistic driving behavior.

First, the challenge in analyzing large data sets, needed to be tackled to prevent the loss of important data characteristics, to apply techniques which change the structure of time series data, and to encounter models that are not interpretable, transparent, and explainable. Therefore, this study applied a time-series abstraction method (SAX, Symbolic Aggregate Approximation) together with topic modeling using LDA, which enables explicate approaches making models and results interpretable, transparent and explainable.

In the experiments of this study, symbolic representations (SAX) were applied, for which then Latent Dirichlet Allocation (LDA) for probabilistic topic modeling was applied. To answer the first question a general LDA-model was created, which covered general information about the driving experiments, which were held for this study. The largest topic in our first (simple) model represented city driving. In this topic a stationary state of the car was over-represented. Especially in urban settings, in which roads are more crowded, and traffic lights are more present, it is more likely to stand still with a vehicle. Besides this state, Topic 1 was represented by speeds, which were a reflection of the speed limits that were present in the urban environment. The second topic, represented highway driving, in which SAX-words with high speed occurrences were over represented.

Furthermore, this study aimed at answering whether subtile behavioral driving topics could be distinguished form naturalistic driving behavior. We analyzed whether more complex topic models would enable more powerful insights for identifying behavioral patterns. Inclusion of $n$-grams led to new insights in the data, as bi- and trigrams provided more information about occurrences in the data which followed each other. More specifically, besides city- and highway driving, this model could provide more detailed informa-

tion about the situations that occurred during the experiment. For example, a clear distinction between high way driving, with low density in traffic was revealed, compared to high density (i.e., rush hour). Also, the model detected driving behavior in which a maximum speed of 70 km/h was allowed, but high density of vehicles and traffic lights were present.

With respect to the first and second research question, this study was interested in detecting differences between alert and unalert participants. To answer this question, a Psychomotor Vigilance Test was included in the experiments. The results of the test were used to label participants, and to measure whether there were differences in topic probabilities between alert and unalert participants. In general, the results indicated that the labeled groups did not differ, as topic probabilities were in similar ranges. More interestingly, topic 1 in the LDA-model including $n$-grams indicated that four alert drivers had a high topic probability for highway driving, which were the only participants in the sample who drove on a highway. However, future research is required to investigate this more as the sample was too small to validate the results.

## 7.1 Limitations

The current study applied topic modeling using LDA to extract clusters from NDD. Although application of this method is validated, the results of the current study need to be interpreted with caution as the study had several limitations. First, this study was restricted to a sample size of 25 participants, of which 30 minutes of NDD was collected on average. Compared to the study of McLaurin et al. (2014), who used a sample size of 10,705 rides, the sample size of this study did not allow to generalize the results. Moreover, the presented topic probabilities between topics and labels in Appendix D proof the limitation of this study as the topic probabilities between groups of participants was unevenly distributed (i.e., 6 unalert versus 18 alert participants). Thus, the output of this study led to results, which need to be interpreted with caution as they cannot be generalized due to the experimental setup.

Second, the experimental setup in the study of McLaurin et al. (2014) included a clear diversified group (i.e., OSA patients versus non-OSA patients). Including these two clear distinguished groups simplified the study of McLaurin et al. (2014) as topic probabilities could be compared between both groups, by comparing the means of the topic probabilities. The current study did not have access to a human subject pool with OSA-patients, and thus, applied PVT to objectively quantify the vigilance of participants to facilitate labelling. Appendix A and C presented the results of PVT. As explained in Section 5.3.2 participants were divided as alert or unalert. More

specifically, this study determined alertness level based on the quantity of true errors and false positives, combined with the distribution of reaction times of correct PVT input. These results should be improved as figures in Appendix C did not differ significantly, due to overlapping box plots. Hence, the labeling of participants based on their PVT results, should be interpreted with caution.

Third, this study applied a conversion of time series data to symbolic representation with SAX. This conversion led to a more intuitive dataset, which makes it possible to interpret NDD by abstraction. Although this method reveals nuances in NDD, other techniques have the potential to cluster sequences of SAX representations. For example, voting experts (VOX), is a more sophisticated algorithm, which is able to redefine continuous strings to words. McDonald et al. (2013) were able to create clusters of strings, presented by SAX-letters, which enabled to analyze part of NDD and to clearly define the situations in NDB. This study, investigated word occurrences in topics, by analyzing SAX representation in LDA models, but VOX-representations could benefit the symbolic representation of data as new defined strings would explain more about occurrences in NDB.

## 7.2 Future research

Despite the promising results of this study, questions remain. Further work is required to establish the viability of PVT in NDB settings. This study aimed at objectively analyzing alertness among participants by applying PVT. Future longitudinal studies, might consider to apply PVT in their experimental setups. There are numerous PVT applications for Android or iOS, which make it affordable and efficient to apply PVT. This makes it possible to connect multiple in-vehicle sensors, combined with inputs of individuals in real driving scenario's. Longitudinal studies would benefit receiving input from participants over several moments in a day, which would make it possible to analyze the vigilant state of participants over a longer period of time, compared to NDB.

The transformation of continuous data to symbolic representations has offered this study to build topic models, which were based speed and acceleration. The occurrences in this corpus were combined by means of bi-, and trigrams in preparation for the LDA-model. As previously described in the limitations section, future research should aim at applying VOX algorithm to a naturalistic driving dataset as it would lead to more useful letter combinations. In text mining, VOX segments collection of strings (i.e., without spaces) by finding the highest entropy. Entropy is the lowest within words,

which is where VOX algorithm finds frequent occurring combinations and segments them. Future research would benefit by applying this algorithm to find even more meaningful symbolic representations, which would create topic models with more insightful and intuitive words. Also, words that are defined with VOX, could be combined with GPS data, to determine the exact situations in real driving situations. However, automated techniques need to be implemented as manual labeling would be time consuming and inefficient.

# 8 Conclusion

The results of the current study, have revealed different behavioral patterns providing novel insights in LDA-analysis. Furthermore, we have shown that the applied methodology using extended LDA models allows to obtain more comprehensive models for identifying behavioral patterns. To the best of our knowledge, this is the first time that such methods have been applied in this context. This contribution might provide the opportunity to detect patterns in naturalistic driving behavior, which would not have been detected with human interference. This way, it is possible to detect pre-accidental situations, which provide more information about the driving behavior of people. Also, if pre-accidental information is available, systems and applications could be developed, which could serve as a warning system, to prevent people from risky driving behavior.

For future work, we aim at analyzing richer behavioral profiles on topic models, utilizing subgroup discovery (Hendrickson, Wang, & Atzmueller, 2018). Then, also appropriate methods for visualizing and detailed inspection are interesting directions to consider. Furthermore, the spatio-temporal analysis of the (abstracted) time-series data using data mining and network analysis (Atzmueller, 2014; Atzmueller & Lemmerich, 2013; Giannotti, Nanni, Pinelli, & Pedreschi, 2007; Verhein & Chawla, 2008) as well as contextualized approaches for local exceptionality modeling and mining, Atzmueller (2016); Atzmueller, Schmidt, and Kibanov (2016); Harri, Filali, and Bonnet (2009) are interesting directions for future research.

# References


Aarts, L., Weijermars, W., Schoon, C., & Wesemann, P. (2008). *Maximaal 500 verkeersdoden in 2020: waarom eigenlijk niet.* Maatregel.

Abouelnaga, Y., Eraqi, H. M., & Moustafa, M. N. (2017). Real-time distracted driver posture classification. *arXiv preprint arXiv:1706.09498*.

Aksan, N., Dawson, J., Tippin, J., Lee, J. D., & Rizzo, M. (2015). Effects of fatigue on real-world driving in diseased and control participants. In *Proceedings of the... international driving symposium on human factors in driver assessment, training, and vehicle design* (Vol. 2015, p. 268).

Arsintescu, L., Kato, K. H., Cravalho, P. F., Feick, N. H., Stone, L. S., & Flynn-Evans, E. E. (2017). Validation of a touchscreen psychomotor vigilance task. *Accident Analysis & Prevention*.

Aryal, A., Ghahramani, A., & Becerik-Gerber, B. (2017). Monitoring fatigue in construction workers using physiological measurements. *Automation in Construction*, *82*, 154–165.

Atzmueller, M. (2014, June). Data Mining on Social Interaction Networks. *Journal of Data Mining and Digital Humanities*, *1*.

Atzmueller, M. (2016). Detecting Community Patterns Capturing Exceptional Link Trails. In *Proc. ieee/acm asonam.* Boston, MA, USA: IEEE Press.

Atzmueller, M., & Lemmerich, F. (2013). Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. *IJWS*, *2*(1/2), 80-112.

Atzmueller, M., Schmidt, A., & Kibanov, M. (2016). DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails. In *Proc. www 2016 (companion).*

Basner, M., & Dinges, D. F. (2011). Maximizing sensitivity of the psychomotor vigilance test (pvt) to sleep loss. *Sleep*, *34*(5), 581–591.

Batool-Anwar, S., Kales, S. N., Patel, S. R., Varvarigou, V., DeYoung, P. N., & Malhotra, A. (2014). Obstructive sleep apnea and psychomotor vigilance task performance. *Nature and science of sleep*, *6*, 65.

Battiato, S., Farinella, G. M., Gallo, G., & Giudice, O. (2018). On-board monitoring system for road traffic safety analysis. *Computers in Industry*, *98*, 208–217.

Belakhdar, I., Kaaniche, W., Djemal, R., & Ouni, B. (2018). Single-channel-based automatic drowsiness detection architecture with a reduced number of eeg features. *Microprocessors and Microsystems*, *58*, 13–23.

Bener, A., Lajunen, T., Özkan, T., Yildirim, E., & Jadaan, K. S. (2017). The impact of aggressive behaviour, sleeping, and fatigue on road traffic crashes as comparison between minibus/van/pick-up and commercial taxi drivers. *Journal of Traffic and Transportation Engineering*, *5*, 21–31.

Bener, A., Yildirim, E., Özkan, T., & Lajunen, T. (2017). Driver sleepiness, fatigue, careless behavior and risk of motor vehicle crash and injury: Population based case and control study. *Journal of Traffic and Transportation Engineering (English Edition)*, *4*(5), 496–502.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bora, D. J., Gupta, D., & Kumar, A. (2014). A comparative study between fuzzy clustering algorithm and hard clustering algorithm. *arXiv preprint arXiv:1404.6059*.

Botzer, A., Musicant, O., & Perry, A. (2017). Driver behavior with a smartphone collision warning application–a field study. *Safety science*, *91*, 361–372.

Brodsky, W., Olivieri, D., & Chekaluk, E. (2018). Music genre induced driver aggression: A case of media delinquency and risk-promoting popular culture. *Music & Science*, *1*, 2059204317743118.

Brunnauer, A., Segmiller, F. M., Löschner, S., Grun, V., Padberg, F., & Palm, U. (2018). The effects of transcranial direct current stimulation (tdcs) on psychomotor and visual perception functions related to driving skills. *Frontiers in behavioral neuroscience*, *12*, 16.

Cantin, V., Lavallière, M., Simoneau, M., & Teasdale, N. (2009). Mental workload when driving in a simulator: Effects of age and driving complexity. *Accident Analysis & Prevention*, *41*(4), 763–771.

CBS. (2018). *In 2017 meer verkeersdoden op de fiets dan in de auto.* Retrieved from `https://www.cbs.nl/nl-nl/nieuws/2018/17/in-2017-meer-verkeersdoden-op-de-fiets-dan-in-de-auto` (Accessed: 2018-09-16)

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

Chen, H.-Y. W., Donmez, B., Hoekstra-Atwood, L., & Marulanda, S. (2016). Self-reported engagement in driver distraction: An application of the theory of planned behaviour. *Transportation research part F: traffic psychology and behaviour*, *38*, 151–163.

Chen, J. H., Goldstein, M. K., Asch, S. M., Mackey, L., & Altman, R. B. (2017). Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *Journal of the American Medical Informatics Association*, *24*(3), 472–480.

Chowdhury, A., Chakravarty, T., Ghose, A., Banerjee, T., & Balamuralidhar, P. (2018). Investigations on driver unique identification from smartphone gps data alone. *Journal of Advanced Transportation*, *2018*.

Constantinescu, Z., Marinoiu, C., & Vladoiu, M. (2010). Driving style analysis using data mining techniques. *International Journal of Computers Communications & Control*, *5*(5), 654–663.

Doig, C. (2015). *Introduction to topic modeling in python.* `http://chdoig.github.io/pytexas2015-topic-modeling/`.

Farrelly, C. M., Schwartz, S. J., Amodeo, A. L., Feaster, D. J., Steinley, D. L., Meca, A., & Picariello, S. (2017). The analysis of bridging constructs with hierarchical clustering methods: An application to identity. *Journal of Research in Personality*, *70*, 93–106.

Garbarino, S., Magnavita, N., Guglielmi, O., Maestri, M., Dini, G., Bersi, F. M., ... Durando, P. (2017). Insomnia is associated with road accidents. further evidence from a study on truck drivers. *PLoS one*, *12*(10), e0187256.

Giannotti, F., Nanni, M., Pinelli, F., & Pedreschi, D. (2007). Trajectory Pattern Mining. In *Proc. sigkdd* (pp. 330–339).

González, J. R. C., Romero, J. J. F., Guerrero, M. G., & Calderón, F. (2015). Multi-class multi-tag classifier system for stackoverflow questions. In *Power, electronics and computing (ropec), 2015 ieee international autumn meeting on* (pp. 1–6).

Harri, J., Filali, F., & Bonnet, C. (2009). Mobility Models for Vehicular Ad Hoc Networks: A Survey and Taxonomy. *IEEE Communications Surveys & Tutorials*, *11*(4).

Hendrickson, A., Wang, J., & Atzmueller, M. (2018). Identifying Exceptional Descriptions of People Using Topic Modeling and Subgroup Discovery. In *Proc. ismis.* Berlin/Heidelberg, Germany: Springer.

Hoffman, M. D., Blei, D. M., & Bach, F. (n.d.). Online learning for latent dirichlet allocation.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, *31*(8), 651–666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264–323.

Jones, M. J., Dunican, I. C., Murray, K., Peeling, P., Dawson, B., Halson, S., ... Eastwood, P. R. (2018). The psychomotor vigilance test: a comparison of different test durations in elite athletes. *Journal of sports sciences*, *36*(18), 2033–2037.

Kumar, V., Dhok, S. B., Tripathi, R., & Tiwari, S. (2014). A review study of hierarchical clustering algorithms for wireless sensor networks. *International Journal of Computer Science Issues (IJCSI)*, *11*(3), 92.

Li, S., Wang, W., Mo, Z., & Zhao, D. (2018). Clustering of naturalistic driving encounters using unsupervised learning. *arXiv preprint*

*arXiv:1802.10214*.

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, *15*(2), 107–144.

Loh, S., Lamond, N., Dorrian, J., Roach, G., & Dawson, D. (2004). The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 339–346.

McDonald, A. D., Lee, J. D., Aksan, N. S., Dawson, J. D., Tippin, J., & Rizzo, M. (2013). The language of driving: Advantages and applications of symbolic data reduction for analysis of naturalistic driving data. *Transportation research record*, *2392*(1), 22–30.

McLaurin, E., McDonald, A. D., Lee, J. D., Aksan, N., Dawson, J., Tippin, J., & Rizzo, M. (2014). Variations on a theme: Topic modeling of naturalistic driving data. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 2107–2111).

Neale, V. L., Dingus, T. A., Klauer, S. G., Sudweeks, J., & Goodman, M. (2005). An overview of the 100-car naturalistic study and findings. *National Highway Traffic Safety Administration, Paper*, *5*, 0400.

Park, D., Lee, M., Park, S. E., Seong, J.-K., & Youn, I. (2018). Determination of optimal heart rate variability features based on svm-recursive feature elimination for cumulative stress monitoring using ecg sensor. *Sensors (Basel, Switzerland)*, *18*(7).

Puschmann, D., Barnaghi, P., & Tafazolli, R. (2018). Using lda to uncover the underlying structures and relations in smart city data streams. *IEEE Systems Journal*, *12*(2), 1755–1766.

Radun, I., Radun, J., Wahde, M., Watling, C. N., & Kecklund, G. (2015). Self-reported circumstances and consequences of driving while sleepy. *Transportation research part F: traffic psychology and behaviour*, *32*, 91–100.

Remmits, Y. (2017). Finding the topics of case law: Latent dirichlet allocation on supreme court decisions.

Saxby, D. J., Matthews, G., & Neubauer, C. (2017). The relationship between cell phone use and management of driver fatigue: it's complicated. *Journal of safety research*, *61*, 129–140.

Sehgal, S., & Kapoor, R. (2018). Mathematical relationship among visual reaction time, age and bmi in healthy adults. *INDIAN JOURNAL OF APPLIED RESEARCH*, *8*(8).

Song, W., Woon, F. L., Doong, A., Persad, C., Tijerina, L., Pandit, P., ... Giordani, B. (2017). Fatigue in younger and older drivers: Effectiveness of an alertness-maintaining task. *Human factors*, *59*(6), 995–1008.

Venkatraman, V., Liang, Y., McLaurin, E. J., Horrey, W. J., & Lesch, M. F. (2017). Exploring driver responses to unexpected and expected events

using probabilistic topic models.

Verhein, F., & Chawla, S. (2008). Mining Spatio-Temporal Patterns in Object Mobility Databases. *Data mining and knowledge discovery*, *16*(1), 5–38.

Virdi, G., & Madan, N. (2018). Review on various enhancements in k means clustering algorithm.

Wohleber, R. W., & Matthews, G. (2016). Multiple facets of overconfidence: Implications for driving safety. *Transportation research part F: traffic psychology and behaviour*, *43*, 265–278.

Yang, L., Li, X., Guan, W., Zhang, H. M., & Fan, L. (2018). Effect of traffic density on drivers' lane change and overtaking maneuvers in freeway situation-a driving simulator based study. *Traffic injury prevention*, 1–25.

Yang, L., Ma, R., Zhang, H. M., Guan, W., & Jiang, S. (2018). Driving behavior recognition using eeg data from a simulated car-following experiment. *Accident Analysis & Prevention*, *116*, 30–40.

Zhang, G., Yau, K. K., Zhang, X., & Li, Y. (2016). Traffic accidents involving fatigue driving and their extent of casualties. *Accident Analysis & Prevention*, *87*, 34–42.

# Appendices

# A PVT Error Results

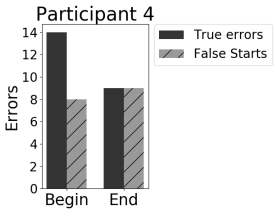

(a)



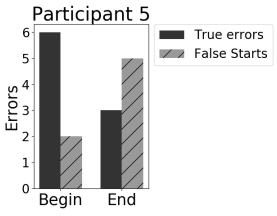

(b)



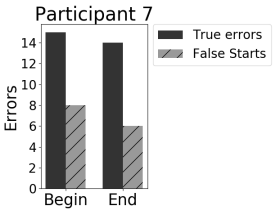

(c)




(d)

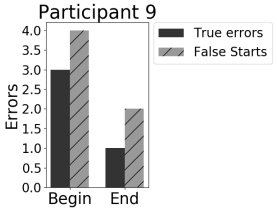



(e)

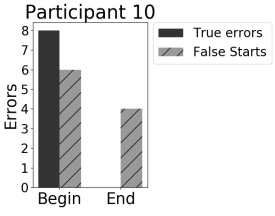



(f)

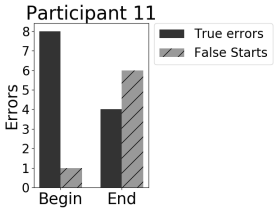

(g)



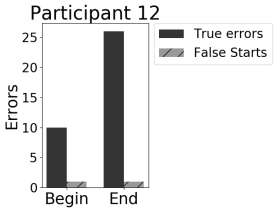

(h)



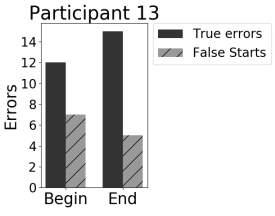

(i)



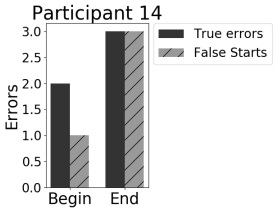

(j)



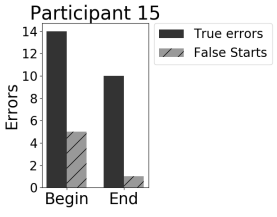

(k)



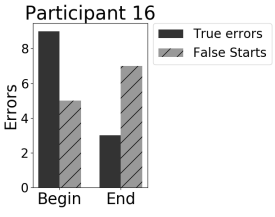

(l)




(m)




(n)




(o)

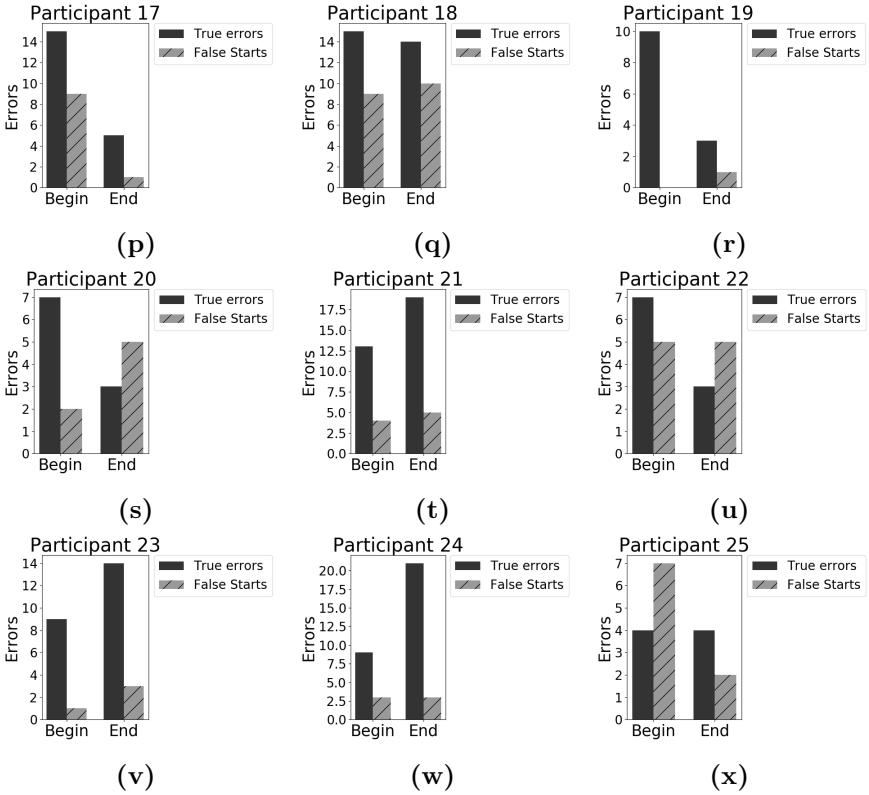


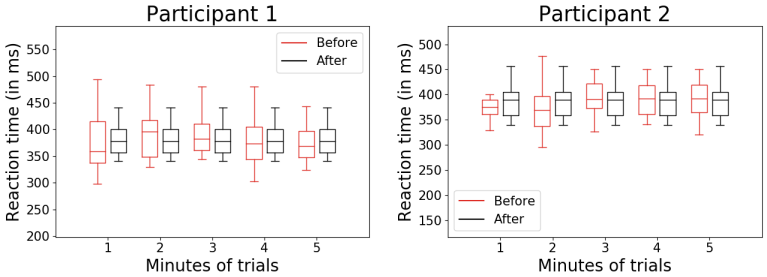

**Figure A.1** Amount of errors per participants distributed over two moments of PVT. First, the errors before the driving experiment are displayed. Second, the errors are displayed after the driving experiment. True errors are defined as reaction times >500 ms. False starts are defined as reaction times <100 ms.

# B Labels of Alertness State

| Participant | Label |
|---|---|
| 1 | Unalert |
| 2 | Alert |
| 3 | Alert |
| 4 | Alert |
| 5 | Alert |
| 7 | Alert |
| 8 | Alert |
| 9 | Alert |
| 10 | Alert |
| 11 | Alert |
| 12 | Unalert |
| 13 | Alert |
| 14 | Unalert |
| 15 | Alert |
| 16 | Alert |
| 17 | Alert |
| 18 | Alert |
| 19 | Alert |
| 20 | Alert |
| 21 | Unalert |
| 22 | Alert |
| 23 | Unalert |
| 24 | Unalert |
| 25 | Alert |

**Table B.1** Labels of alertness state after conducted driving experiment. The labels are based on the increased or decreased number of errors (i.e., False Starts and True Error). In total, 6 participants were denoted as Unalert, and 18 participants as Alert.

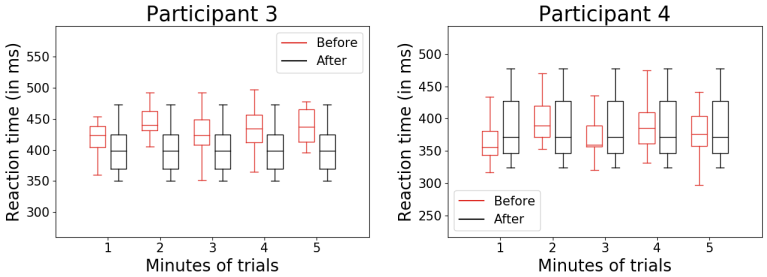
# C PVT Reaction Time Results




(a)



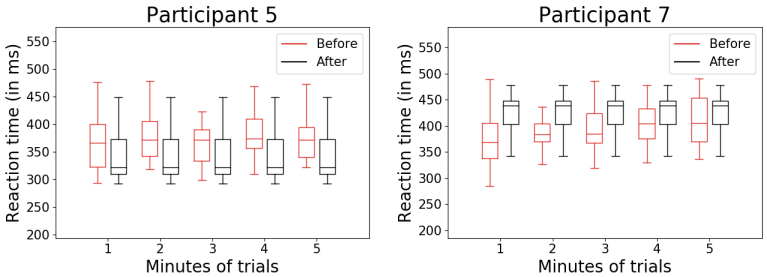

(b)




(c)



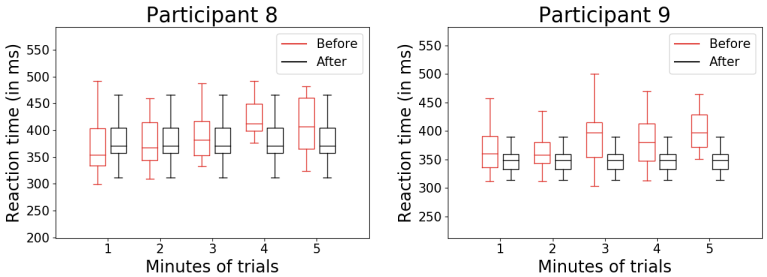

(d)




(e)



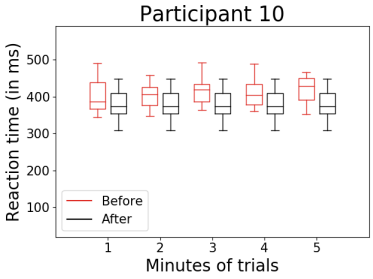

(f)



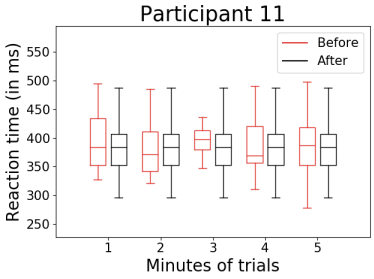

(g)



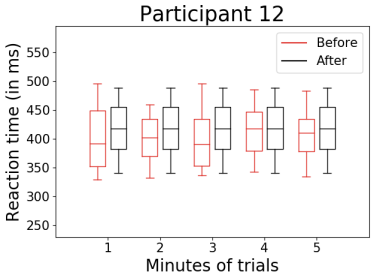

(h)


Participant 10
Reaction time (in ms)
Minutes of trials
Before
After

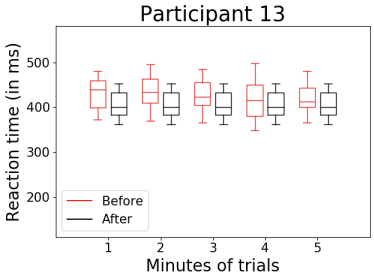
(i)


Participant 11
Reaction time (in ms)
Minutes of trials
Before
After

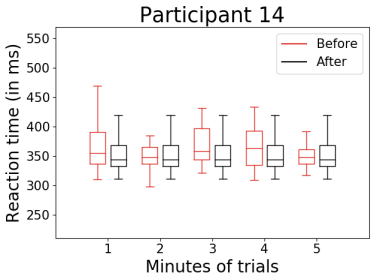
(j)


Participant 12
Reaction time (in ms)
Minutes of trials
Before
After

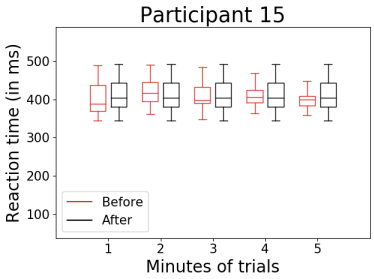
(k)


Participant 13
Reaction time (in ms)
Minutes of trials
Before
After

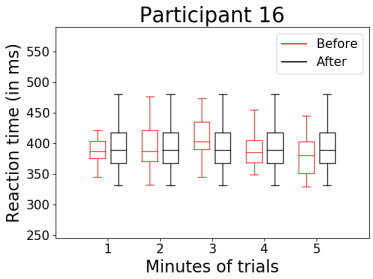
(l)


Participant 14
Reaction time (in ms)
Minutes of trials
Before
After

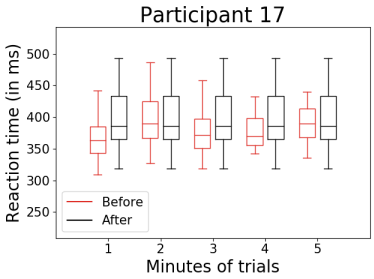
(m)


Participant 15
Reaction time (in ms)
Minutes of trials
Before
After

(n)


Participant 16
Reaction time (in ms)
Minutes of trials
Before
After

(o)


Participant 17
Reaction time (in ms)
Minutes of trials
Before
After

(p)

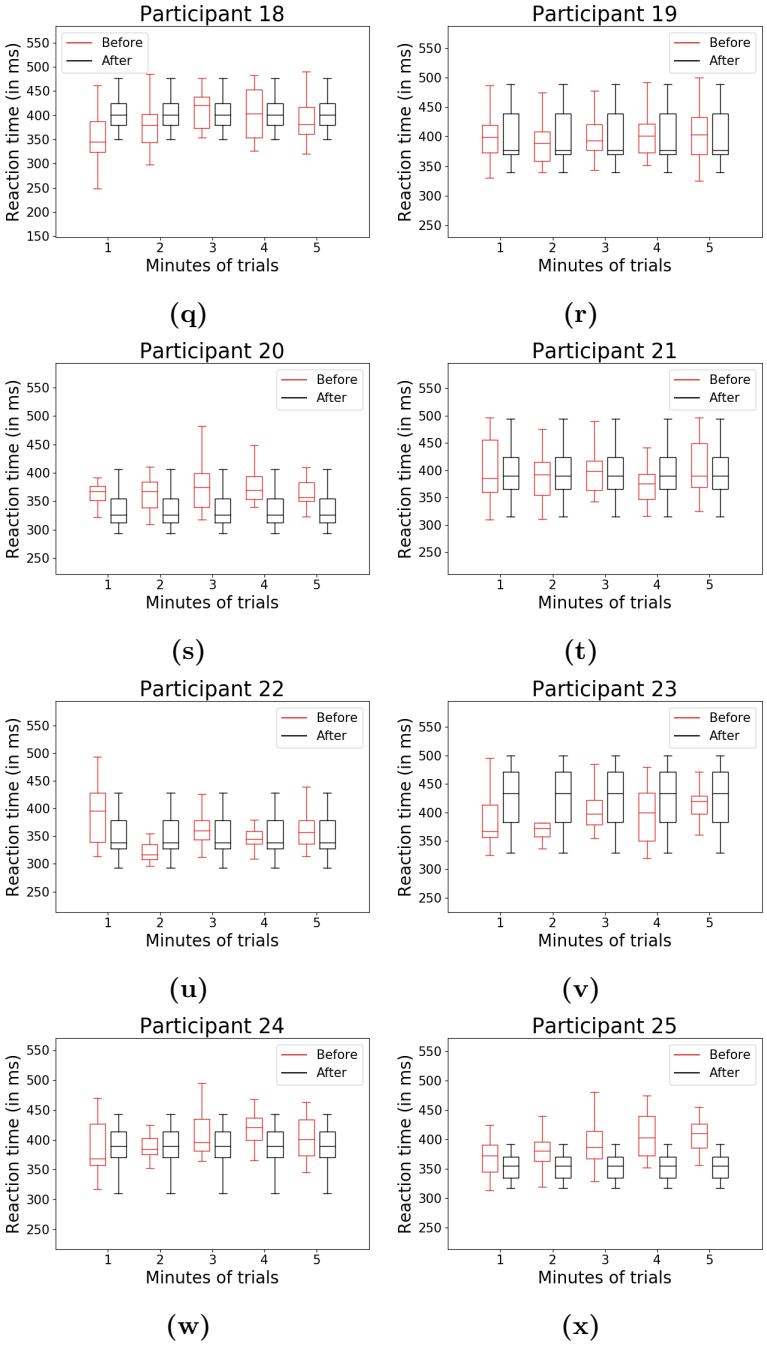



**Figure C.1** Psychomotor vigilance test reaction time results. Each test was conducted before and after the driving experiment. Pre- and after-results for each 5-minute PVT, are presented aside, with whiskers presenting the standard deviation. The means are presented by the horizontal lines in each boxplot.

# D Topic probability of alertness state.

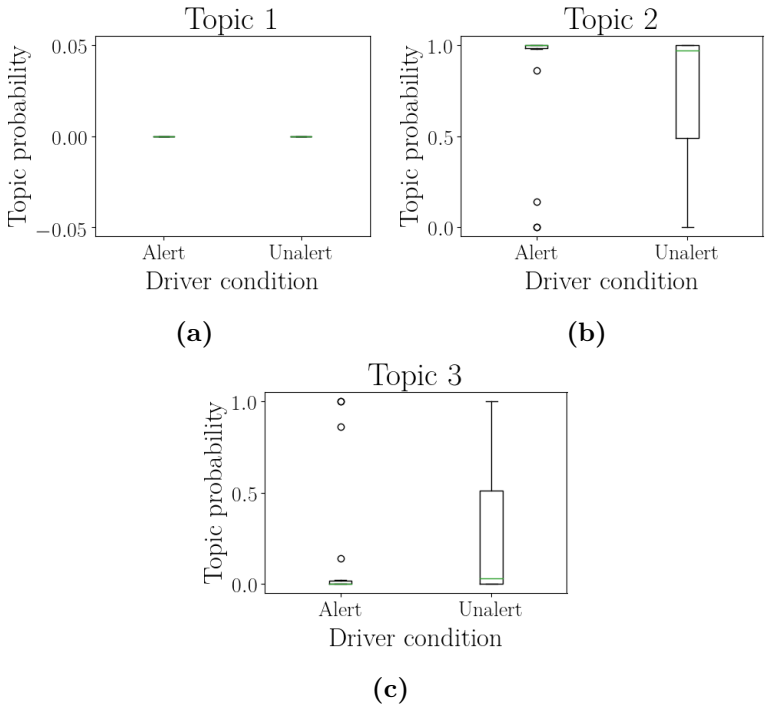



**Figure D.1** Mean topic probabilities of alert versus unalert participants. The standard deviations are shown as error bars. The boxplots include the results of the standard LDA-model.

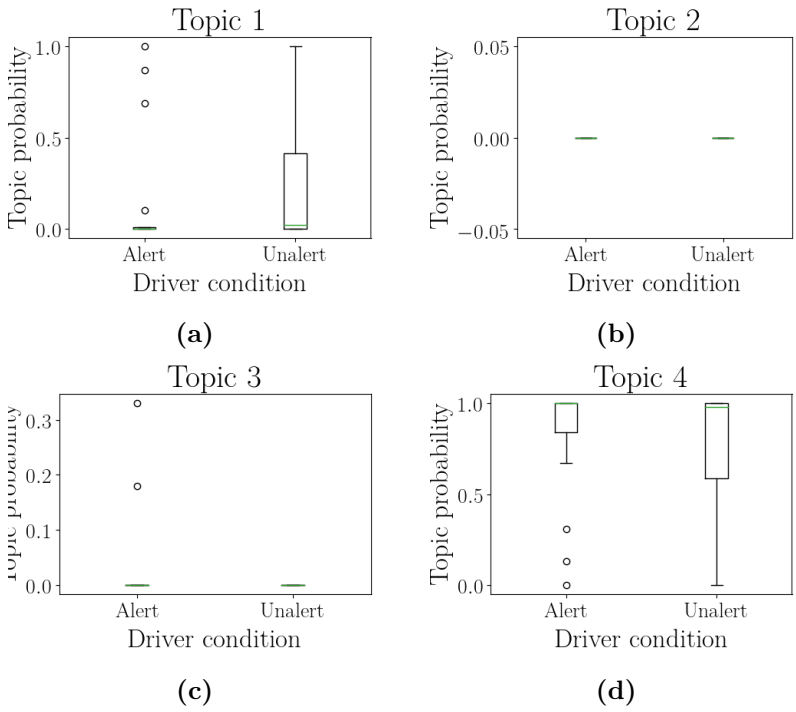



**Figure D.2** Mean topic probabilities of alert versus unalert participants. The standard deviations are shown as error bars. The boxplots include the results of the LDA-model, inluding bi-and trigrams.

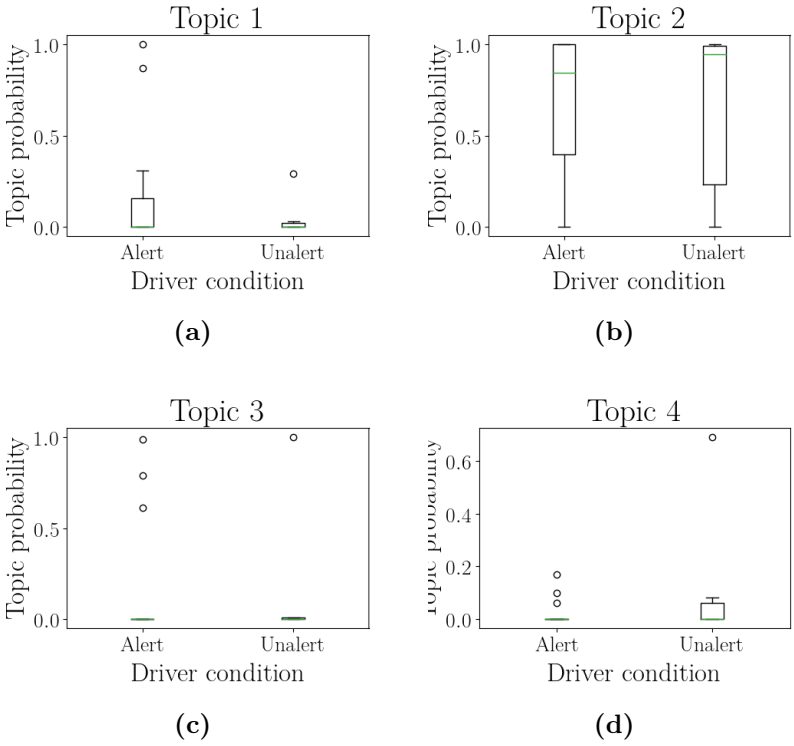



**Figure D.3** Mean topic probabilities of alert versus unalert participants. The standard deviations are shown as error bars. The boxplots include the results of the LDA-model, inluding bi-and trigrams and max-features = 100.

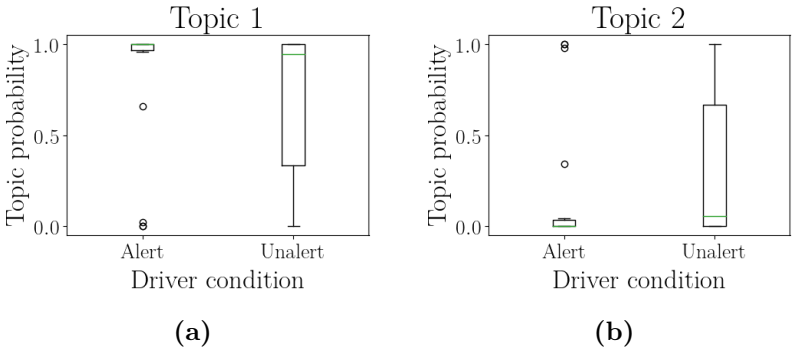



**Figure D.4** Mean topic probabilities of alert versus unalert participants. The standard deviations are shown as error bars. The boxplots include the results of the LDA-model, inluding bi-and trigrams and min-df = 0.1, and max-df= 0.7.