Behavioral Topic Modeling on Naturalistic Driving Data

Sebastiaan Merino and Martin Atzmueller

Tilburg University, The Netherlands
{s.m.merino-norambuena, m.atzmuller}@uvt.nl

Abstract. Identifying risky driving behavior is of central importance for increasing traffic safety. This paper tackles the task of analyzing real (naturalistic) driving data captured by in-vehicle sensors using interpretable data science methods. In particular, we focus on symbolic time-series abstraction and the subsequent behavioral profile identification using topic modeling approaches. For our experiments, we utilize a real-world dataset. Our results indicate interesting behavioral driving profiles including important patterns and factors for traffic safety modeling.

1 Introduction

An increase in road accidents has become a central issue for researchers to identify risky driving behavior, which increases the chance of road fatalities. To prevent road fatalities from increasing excessively, the Dutch government has the objective to keep the number of fatalities by 2020 under 500 per year [1]. 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, e. g., [17,21,29].

In this context, this paper aims to contribute to improvements of road safety by constructing topic models which describe naturalistic driving behavior. The topic models are constructed using time-series of in-vehicle sensor data. By combining interpretable symbolic abstraction methods with appropriate (topic) modeling approaches, interesting driver profiles, as well as interesting behavioral patterns on driving data can be detected. Our contributions can be summarized as follows: (1) We apply a time-series abstraction method (SAX, Symbolic Aggregate Approximation) that enables interpretable elements to be used in the topic models, ans (2) we investigate more complex behavioural patterns; (3) with that, we propose an interpretable analysis approach enhancing transparency and explainability in the scope of *explicative data mining* [4,5]. For our experiments, we utilize a dataset collected using in-vehicle sensors on naturalistic driving data. Our results indicate interesting driving profiles, as well as behavioral driving patterns which are interpretable given the symbolic data representation.

The rest of the paper is structured as follows. Section 2 discusses related work. After that, Section 3 details the applied methods. Next, Section 4 presents and discusses our results. Finally, Section 6 concludes with a summary and interesting directions for future work.

Preprint of: Sebastiaan Merino and Martin Atzmueller (2018)
Behavioral Topic Modeling on Naturalistic Driving Data.
In: Proc. Benelux Conference on Artificial Intelligence (BNAIC),
Jheronimus Academy of Data Science (JADS), 's-Hertogenbosch, The Netherlands

2 Related Work

Extensive research has already provided many insights in the field of road safety. For example, [28] mentioned in their study that an increase in motorization has lead to "severe traffic-related causalities" [28, p. 34]. In January 2018, the amount of vehicles in The Netherlands increased with 2% to 12.5 million vehicles. As density in traffic increases, traffic safety is influenced negatively as it leads to an increase in lane changing and overtaking cars [27]. More specifically, [12] mentioned lane changing and overtaking cars has negative effects on traffic safety.

2.1 Profiling Driving Behavior

Besides traffic density having an influence on traffic safety, research has also been conducted to study driving behavior [8, 14, 15]. Studies included self-reports of large populations to study the behavior of participants while driving a car. The focus in these 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, [26] concluded 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" [26, 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 him- or herself 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 makes individuals not take precautions such as resting, hands-free driving, or not making use of a mobile phone while driving [23].

2.2 Naturalistic Driving Data Abstraction and Modeling

In a study conducted by [21], the authors highlighted the importance of naturalistic driving data (NDD) in studying driving behavior. However, the authors mentioned NDD-analyis is challenging as "the large amount of data commonly collected during naturalistic driving studies makes comprehensive analysis prohibitive without some type of data reduction" [21, p. 2107]. A possible consequence of this reduction might lead to omit subtle information or even the broader context of a data set. Moreover, a challenge to analyze large amounts of data is the interpretation of continuous data. Therefore, [21] included a symbolic representation of time series data by applying Symbolic Aggregate Approximation (SAX) [20], which transparently retains the data characteristics. The conversion of time series data is executed by transforming continuous instances to alphabetical representations. Time series output is normalized and divided into equal sized ranges. Then, each range is presented by a letter, so instances which coincide in a range are assigned to a letter. By applying this transformation to time series, with multiple variables, a *bag-of-words*-model is created. The method

of [21] allowed consecutive studies not only to analyze large amounts of time series data, but also to combine various variables [21, 22, 24]: in the study of [22], traffic data was combined with weather data, while [24] were able to analyze expected events during driving situations of individuals.

The SAX representation of time series data enables extended analysis: More specifically, Probabilistic Topic Modelling (PTM) [9] makes it possible to apply an unsupervised learning exploration of the data. Intuitively, the output of PTM is a topic model which captures occurrence distributions 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 and find subtle patterns. [24], for example, measured differences in expected and unexpected events involving crosswinds, where they 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 continuous variables to a verbal representation, which was necessary as they applied Latent Dirichlet Allocation (LDA) [9,10] for probabilistic topic modeling in their study.

3 Methods

This section establishes the methodology of the current study. We first summarize the collected dataset and the applied preprocessing, before we present the modeling method in detail.

3.1 Data collection

Naturalistic driving data for this study were collected by conducting twenty-five real-life experiments. In collaboration with Crossyn Automotive B.V.[1] in the Netherlands, this study made use of an in-vehicle sensor, which recorded rides during experiments providing GPS sensor data. The GPS data made it possible to measure linear acceleration and speed. 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)[2]. Attendees were invited to participate by signing a consent form, data collection was then anonymized. Participants were instructed that they would receive oral directions throughout the driving task, meaning they did not have to navigate 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.[3] Finally, participants were allowed to have a casual conversation during the experiment in order to imitate naturalistic driving behavior. Most of the routes showed similarities, covering approximately 30 minutes of driving on long uneventful roads, where the speed limits vary from 50 km/h to 70 km/h.

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

[2] M = mean, SD = standard deviation.

[3] In their study, [11] concluded that hostile music can lead to distracted drivers.

3.2 Preprocessing and Feature Extraction

The final dataset consists of 24 rides: One experiment/ride was excluded from the dataset, as the in-vehicle sensor lost its connection due to an unknown reason. 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).

The data processing using SAX on the collected time series data was achieved by creating segments of the value range of the respective time series in the data, in which each segment was represented by an alphabetical letter. Then, the alphabet of SAX letters needed to be determined. Referring to the study of [21] in which nine groups were defined, this study applied a similar amount of symbols. However, after conversion of speed to SAX-letters, eight letters could be generated at most. Therefore, the normalized data were divided into eight scales for the attribute speed and nine for the attribute (linear) acceleration. Finally, the letters in the alphabet were represented by nine unique letters. Table 1 illustrates the alphabetical representation of letters used for this study, and the corresponding value ranges that these SAX-letters represented.

	Range			
SAX letters	Speed (km/h)	N	Acceleration (g)	N
A/a	0.0 to 1.0	10566	−1.1890 to 0.0760	3728
B/b	2.0 to 16.0	4161	−0.0708 to −0.0472	2474
C/c	17.0 to 28.0	4407	−0.0437 to −0.0283	6224
D/d	28.7 to 38.0	5417	−0.0212 to −0.0094	56
E/e	39.0 to 48.1	10513	−0.0081 to 0.0071	24230
F/f	48.4 to 59.3	4011	0.0089 to 0.0239	62
G/g	60.0 to 74.0	3054	0.0283 to 0.0438	5208
H/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 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 uppercase letter (i.e. from "A" to "H") which coincided with speed. Subsequently, the second (lowercase) 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 shaped based on the combinations of speed, and acceleration letters of the dataset.

The current study exhibits similarities with the frequency of speed letters in the study of [21], partially reproducing their results, but in another context, since we aim at identifying distinctive behavioral profiles in a general setting. As described above, for readability purposes, letters which present speed intervals

are mentioned by an uppercase (capital) letter and acceleration intervals by lowercase letters. Table 1 already indicated that the most occurring letters in the data set were "A" ($N = 10556$), "E" ($N = 10513$), and "H" ($N = 6669$). As "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 [21]. Followed by the "E", the letters "F", and "G" range between speeds of 48.4 to 74.0 km/h, which represents roads where the speed limit is 70 km/h. Henceforth, 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 [21], in which the letter "H" was defined as high speed driving. Therefore, the current pre-processing steps indicate similarities to this previous study.

Whereas probabilistic topic modelling applies text documents in order to conduct its analysis, this study makes use of driving data, which typically consists of continuous variables. Therefore, speed and acceleration were first converted to SAX words. Then, since LDA required a document word input for topic modeling, the collection of words in documents was transformed into a *bag-of-words* (BOW) representation, a standard format which is applied in natural language processing, containing the occurring words and their frequencies.

In order to obtain the LDA model, it is required to maximize the log-likelihood [10]. We performed grid search, to determine the topic model with the best fitting topics empirically. The used parameters were: number of topics and learning decay. LDA requests for input for the amount of topics. Therefore, a range from 1 to 5 was given as input. Learning decay is applied to control the learning rate. The values which are 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. The grid search indicated that the best results were obtained with a learning decay of 0.5, yielding three topics.

Unlike the study of [21], in which no words were excluded from the data set, this study included also pre-processing steps which are typically applied in text mining analyses. The following further steps were included in the analysis, as detailed below in Section 4:

- n-grams: For the analysis bi-, and trigrams were included.
- n-grams/max-features: This parameter controls the maximum amount of features (e. g., only 100 words) to include in the BOW.
- Selective n-grams: Words which appeared only in one document, or which appeared in more than 50% of the documents, were excluded.

4 Results

In this section, we first focus on probabilistic topic modeling using LDA, reproducing similar results as presented in [21]. After that, we present results on more complex topic models enabling a more comprehensive analysis on complex driving patterns.

4.1 Topic Model Description

The fitted topic model 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[4].

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 2 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 2: Topics of top five SAX words with highest weights in descending order.

Topic 1: City Driving Topic 1 has the strongest weight for word "A_e", which is characterized by the letter "A" for speed and "e" for acceleration. Previously, Table 1 described the ranges in which each letter coincided, meaning that "A_e" indicates the car in a stationary position. The second strongest word (i.e., "E_e") identifies city driving while accelerating constantly. Third, the letter "H_e" gives, defines a constant higher speed during a ride (i.e., 75.0 and 124.0 km/h) with no acceleration. The fourth word in topic 1 is "E_g", which describes a city driving speed with a higher acceleration. Finally, the word "G_e" describes a higher speed (i.e., 60.0 to 74.0 km/h) and again with a no acceleration.

Except for word 4, the top occurring words describe a constant speed between 34.0 and 124.0 km/h. Speed "H" indicates a higher speed, which would occur on high ways. At the same time, the lower bound of this range indicates 75 km/h, meaning that roads with speed limits of 70 km/h could occur during a ride. Besides that, the word "A_e" is a frequent occurring word in the dataset, and is listed on top of the other words, indicating that modest speeds are included in this topic, including stopping motions (i.e., stationary position). Thus, topic 1 is described as city driving.

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

Topic 2: Complete City Driving Topic 2 is somewhat similar compared to Topic 1. However, discrepancies are found by the presence of 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.

As Topic 2 differs by the two most occurring words, it resembles city driving similarly to Topic 1. However, the top five 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 was included during the experiment was mostly 50 km/h. This word is essential in the definition of city driving resulting in a more complete description.

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. Then, Topic 3 includes word "A_e", which is a stationary representation 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.

Topic Representation in Documents An essential part of LDA analysis is to distinguish which topics belong to which documents. In order to give insights which topic was most dominant in each document, a topic distribution is displayed in Table 3. A surprising result is that 5 rides are clustered as highway driving, which corresponds 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 3: Distribution of dominant topics in current data set.

Table 4 provides a detailed view on all documents which were included in this study. Then, weights of topics indicate to which extent a topic is present in each document. The weights of all topics, which are included in one document, accumulate to 1.0.

In table 4 Topic 2 and 3 are included, while Topic 1 is 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 4.1 highlighted two topics, which in essence appeared to be similar. The results in Table 4 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: Weights of topic occurrences per ride (corresponding to the document number) for Topic 2 and 3. Topic 1 is excluded from the overview as no occurrences were present for this topic in the experimental sample.

4.2 Visualization of 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 1 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 the bubbles are to each other, the more semantic resemblance they share. Figure 1 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.

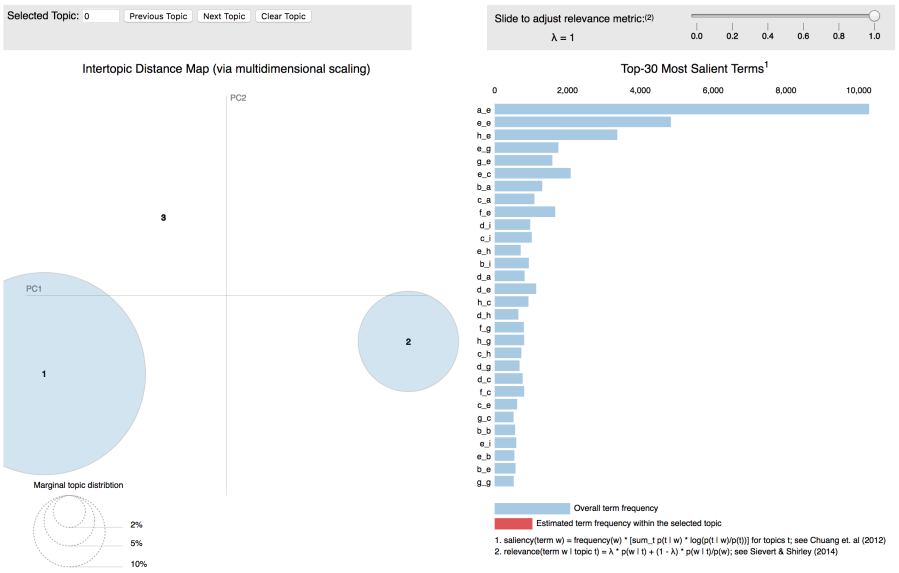

Fig. 1: The topics from the LDA analysis are visualized by applying pyLDAvis. Topics are represented by bubbles on the left side indicating their sizes and respective distances to each other as obtained by multi-dimensional scaling. The right side of the figure displays the word term space, visualizing the respective term frequencies. The words are ranked in descending order of importance.

On the right side of Figure 1 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 pane adapt to the prevalence of the word inside the topic, meaning that the higher the importance of a word in a topic, the larger the size of the bubble.

4.3 Identifying Behavioral Patterns Using Topic Modeling

In the following, we outline the steps for identifying behavioral driving patterns using a more complex topic modeling approach. As we have discussed in Section 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.

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 log-likelihood was achieved with four topics. An overview of each topic, with most occurring vocabulary is shown in Table 5 In order to investigate if the results of this LDA-model differ from the LDA-model described in Section 4.1, we assess each topic individually:

- **Topic 1 (city driving)** forms the largest topic in the LDA-model. The most relevant words indicate similarities with the previously established LDA-model in Section 4.1. More specifically, the top bi-grams, which are shown in Figure 5, 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 1 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 1. "E_e" was an indication of a constant moderate speed (39.0 to 48.1 km/h). Compared to Section 4.1 topic 1 can be determined as city driving as speeds do not exceed speed letter "H", which is higher than 75.0 km/h.
- **Topic 2 (highway driving)** distinguishes itself, compared to Topic 1, by higher speeds. Section 4.1 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 2. More specifically, the highest proportion of "H_e" resides this topic. Thus, this LDA-model has defined Topic 2 as highway driving.
- **Topic 3 (city driving with subtile changes)** shows most resemblance with Topic 1 as speeds do not exceed the letter "H" ($>$ 75.0 km/h). However, the word frequencies in Topic 3 are significantly lower compared to Topic 1. However, a more 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 2 is shown with the 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, which was previously not indicated in Topic 1.
- **Topic 4 (high way driving during rush hour)** is the smallest topic in this LDA-model, as the size in the visualization is minimal. The speed in Topic 4 ranges from minimal (i.e., "A") to a higher speed ("H"). Overall, term frequency is very low in Topic 4, indicating the size of this topic is very small. More interestingly, Topic 4 consists of a combination of high speed,

and a stationary position. Referring to the driving experiments of this study, one participant was subject to rush hour, while driving on a highway. As one participant was subject to this situation, the size of Topic 4 is explained.

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

Table 5: 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 quite informative, we investigated n-grams in more detail. As a first step, we restricted the n-gram set to the set of the top 100 (max-features) in the analysis. This created an optimal LDA-model of four 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 (city driving):** Both aforementioned LDA-models did both include one topic which represented the majority of the data. In this LDA-model, Topic 1 is repeatedly overrepresented. This means, that in the current topic, city driving is represented. All bi- and trigrams, which are included, do not exceed 75.0 km/h. Another observation is that SAX words, which are combined in bi- and trigrams are similar to one another. This indicates that the corpus contains series of driving data which are similar to each other.
- **Topic 2 (high way driving):** Similar to the previous LDA-model, this model included highway driving.
- **Topic 3 (city driving with max 70 km/h)** has the largest representation of "A_e" SAX words. More interestingly, words such as "F_e" and "G_e" are included which are a representation of speeds higher than 60.0 km/h. This result is an indication of events during the driving experiments where participants were driving on a road with a maximum speed of 70 km/h. The inclusion of stationary words (i.e., "A_e") explain traffic lights which participants encountered on this specific part of the road.

– **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 to one participant, who drove on a highway during rush hour in the afternoon.

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. We applied two parameters, i. e., *min-df* and *max-df* for including or excluding features that are underrepresented or overrepresented in the corpus, respectively. 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 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.
– **Topic 2 (high and low acceleration and deceleration):** presents a wider variety of SAX-representations. For example, the first most relevant word in this topic is defined is "B_c B_c B_c", meaning an occurrence during a ride in which a driver would decelerate strongly, while driving a low speed. Typically, in real driving behavior, a state of "B_c" would occur right before the car would reside in stationary position (i.e., "A_e"). The second, most important term in Topic 2 is "C_g C_g C_g". A close look at Table 1 reveals that this trigram explains a state in driving behavior in which the car is strongly accelerating, but still in a fairly low speed. The following terms in Topic 2, are indications of high acceleration and deceleration. For example, the bigram "E_c D_b" is an occurrence in which the vehicle is decelerating. On the contrary, "D_h D_i E_i" indicates a strong acceleration from speed "D" to "E". 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 we removed 50% of most occurring bi- and trigrams in the corpus, this time topics were shaped, which include less situations that occurred during the driving experiments.

5 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 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. On the other hand, the challenge in analyzing large data sets, is 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 explicative approaches making models and results interpretable, transparent and explainable.

In our experiments, we applied symbolic representations (SAX), for which we then applied Latent Dirichlet Allocation (LDA) for probabilistic topic modeling. The topics covered general information about the driving experiments, which were held for this study. The largest topic in our first (simple) model represented city driving, as a state in which the vehicle was stationary 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, 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 information 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.

Compared to the dataset of the study of [21], this study was limited to a sample size of 24 participants, which equalized 24 rides, compared to approximately 10000 rides. A larger sample size would benefit from more nuances in driving data, which would be translated in a potentially broader topic model, with risky driving behavior [13]. Extending the dataset, and also integrating other datasets is aimed for a follow-up study.

6 Conclusions

The results of the current study, have provided 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 systems, to prevent people from risky driving behavior.

For future work, we aim at analyzing richer behavioral profiles on topic models, utilizing subgroup discovery [19]. 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 [2, 6, 16, 25] as well as contextualized approaches for local exceptionality modeling and mining, e. g., [3,7,18] are interesting directions for future research.

References

1. Aarts, L., Weijermars, W., Schoon, C., Wesemann, P.: Maximaal 500 verkeersdoden in 2020: waarom eigenlijk niet (2008)
2. Atzmueller, M.: Data Mining on Social Interaction Networks. Journal of Data Mining and Digital Humanities **1** (June 2014)
3. Atzmueller, M.: Detecting Community Patterns Capturing Exceptional Link Trails. In: Proc. IEEE/ACM ASONAM. IEEE Press, Boston, MA, USA (2016)
4. Atzmueller, M.: Onto Explicative Data Mining: Exploratory, Interpretable and Explainable Analysis. In: Proc. Dutch-Belgian Database Day. TU Eindhoven (2017)
5. Atzmueller, M.: Declarative Aspects in Explicative Data Mining for Computational Sensemaking. In: Proc. DECLARE. Springer, Heidelberg, Germany (2018)
6. Atzmueller, M., Lemmerich, F.: Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. IJWS **2**(1/2), 80–112 (2013)
7. Atzmueller, M., Schmidt, A., Kibanov, M.: DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails. In: Proc. WWW 2016 (Companion). IW3C2 / ACM (2016)
8. Bener, A., Lajunen, T., Özkan, T., Yildirim, E., Jadaan, K.S.: 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 (2017)
9. Blei, D.M.: Probabilistic topic models. CACM **55**(4), 77–84 (2012)
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
11. Brodsky, W., Olivieri, D., Chekaluk, E.: Music Genre Induced Driver Aggression: A Case of Media Delinquency and Risk-Promoting Popular Culture. Music & Science **1**, 2059204317743118 (2018)

12. Cantin, V., Lavallière, M., Simoneau, M., Teasdale, N.: Mental Workload When Driving in a Simulator: Effects of Age and Driving Complexity. Accident Analysis & Prevention **41**(4), 763–771 (2009)
13. Chen, C.P., Zhang, C.Y.: Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. Information Sciences **275**, 314–347 (2014)
14. Chen, H.Y.W., Donmez, B., Hoekstra-Atwood, L., Marulanda, S.: 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 (2016)
15. Garbarino, S., Magnavita, N., Guglielmi, O., Maestri, M., Dini, G., Bersi, F.M., Toletone, A., Chiorri, C., Durando, P.: Insomnia is Associated with Road Accidents. Further Evidence from a Study on Truck Drivers. PLoS one **12**(10), e0187256 (2017)
16. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory Pattern Mining. In: Proc. SIGKDD. pp. 330–339. ACM (2007)
17. Guo, F., Fang, Y.: Individual driver risk assessment using naturalistic driving data. Accident Analysis & Prevention **61**, 3–9 (2013)
18. Harri, J., Filali, F., Bonnet, C.: Mobility Models for Vehicular Ad Hoc Networks: A Survey and Taxonomy. IEEE Communications Surveys & Tutorials **11**(4) (2009)
19. Hendrickson, A., Wang, J., Atzmueller, M.: Identifying Exceptional Descriptions of People Using Topic Modeling and Subgroup Discovery. In: Proc. ISMIS. LNCS, Springer, Berlin/Heidelberg, Germany (2018)
20. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: A Novel Symbolic Representation of Time Series. DMKD **15**(2), 107–144 (2007)
21. McLaurin, E., McDonald, A.D., Lee, J.D., Aksan, N., Dawson, J., Tippin, J., Rizzo, M.: Variations on a Theme: Topic Modeling of Naturalistic Driving Data. In: Proc. Human Factors and Ergonomics Society Annual Meeting. pp. 2107–2111 (2014)
22. Puschmann, D., Barnaghi, P., Tafazolli, R.: Using LDA to Uncover the Underlying Structures and Relations in Smart City Data Streams. IEEE Systems Journal **12**(2), 1755–1766 (2018)
23. Saxby, D.J., Matthews, G., Neubauer, C.: The Relationship between Cell Phone Use and Management of Driver Fatigue: It's Complicated. Journal of safety research **61**, 129–140 (2017)
24. Venkatraman, V., Liang, Y., McLaurin, E.J., Horrey, W.J., Lesch, M.F.: Exploring Driver Responses to Unexpected and Expected Events using Probabilistic Topic Models. In: Proc. International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. pp. 375–381. University of Iowa (2017)
25. Verhein, F., Chawla, S.: Mining Spatio-Temporal Patterns in Object Mobility Databases. Data mining and knowledge discovery **16**(1), 5–38 (2008)
26. Wohleber, R.W., Matthews, G.: Multiple Facets of Overconfidence: Implications for Driving Safety. Transportation research part F: traffic psychology and behaviour **43**, 265–278 (2016)
27. Yang, L., Li, X., Guan, W., Zhang, H.M., Fan, L.: Effect of Traffic Density on Drivers' Lane Change and Overtaking Manoeuvres in Freeway Situation: A Driving Simulator Based Study. Traffic injury prevention pp. 1–25 (2018)
28. Zhang, G., Yau, K.K., Zhang, X., Li, Y.: Traffic Accidents Involving Fatigue Driving and Their Extent of Casualties. Accident Analysis & Prevention **87**, 34–42 (2016)
29. Zheng, Y., Wang, J., Li, X., Yu, C., Kodaka, K., Li, K.: Driving Risk Assessment Using Cluster Analysis Based on Naturalistic Driving Data. In: Proc. International Conference on Intelligent Transportation Systems. pp. 2584–2589. IEEE (2014)