Transportation Research Part A 181 (2024) 103986

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Understanding the expectations of parents regarding their children's school commuting by public transport using latent Dirichlet Allocation

Mariza Motta Queiroz [a,*], Carlos Roque [b], Filipe Moura [a], João Marôco [c]

[a] CERIS, Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
[b] Laboratório Nacional de Engenharia Civil Departamento de Transportes, Núcleo de Planeamento, Tráfego e Segurança, Av do Brasil 101, 1700-066 Lisboa, Portugal
[c] ISPA - Instituto Universitário, R. Jardim do Tabaco 34, 1100-304 Lisboa, Lisbon, Portugal

ARTICLE INFO

Keywords:
Sustainable school commuting
Open-ended survey responses
Text mining
Topic modeling
Latent Dirichlet Allocation

ABSTRACT

Parents' perceptions regarding public transport and active modes influence the youth's acceptance and support for sustainable school commuting. Urban mobility surveys can gather such insights by utilizing closed and open-ended questions. The latter, particularly, holds the potential for nuanced expectations and insights from Public Transport (PT) users, often absent in closed-ended responses. This paper proposes a methodology utilizing Latent Dirichlet Allocation (LDA) to extract valuable information from open-ended survey responses, shedding light on parents' expectations regarding their children's school commute via PT. Analyzing responses from two surveys involving 448 households, with a focus on parents in the Lisbon Metro Area, spanning the school years of 2017–2018 and 2018–2019, and pre-and post-field interventions, our study employs LDA to assess households' criticisms and recommendations for improving public transport services. Our findings illustrate a shift from general criticisms in the initial survey to proactive suggestions in the subsequent one, aligning with marketing efforts to foster more sustainable school commuting with PT. Empirically, our study underscores LDA's efficacy in capturing users' feedback often neglected by closed-ended questions. Effective preprocessing of textual data facilitates streamlined field interventions. Overall, our contribution provides user-centered insights to inform PT policymakers, promoting the incorporation of user-driven enhancements.

1. Introduction

There is a worldwide consensus about the current unsustainable mobility patterns and the need for solutions to respond to and anticipate current and future urban mobility challenges (United Nations, 2016; ITF, 2017). Improving urban mobility, specifically Public Transport (PT), requires intervention strategies that embrace society's travel in general and younger generations in particular. These can expectably have shorter and longer-term impacts on the sustainability of health-related urban mobility patterns (Karanasiou et al., 2014; Buka et al., 2006; Basington, 2008) and are linked to behavior (Long et al., 2019; Baslington, 2008). Short-term impacts

* Corresponding author.
E-mail addresses: marizaqueiroz@tecnico.ulisboa.pt (M. Motta Queiroz), croque@lnec.pt (C. Roque), fmoura@tecnico.ulisboa.pt (F. Moura), jpmaroco@ispa.pt (J. Marôco).

https://doi.org/10.1016/j.tra.2024.103986
Received 7 July 2021; Received in revised form 15 December 2023; Accepted 20 January 2024
Available online 29 January 2024

concern the upfront changes in the mobility of parents and children and, in the longer run, the possible implications for mobility behavior later in children's adult lives, thus contributing to long-lasting societal changes.

Users of urban mobility systems interface with many products and services during their journey and the transport systems are designed to operate and offer seamless experiences (Preston, 2012). Thus, it is crucial to understand and anticipate users' needs in an increasingly consumer-oriented economy and design appropriate and integrated mobility services. Notably, in younger generations, we need to identify the barriers, beliefs, and motivators that can push students to commute to and from school with PT, as the literature lacks information in this particular case (Heelan et al., 2009; Zhu and Lee, 2009; Esztergár-Kiss and Tettamanti, 2019). Accordingly, pushing PT school commuting requires interventions and actions that potentially trigger change. The parents' perception of PT is critical to accepting and incentivizing such modal shifts by youth and youngsters (Carver et al., 2013; Pont et al., 2011; Panter et al., 2008).

Urban mobility surveys can gather such insights by utilizing closed and open-ended questions. The latter, particularly, holds the potential for nuanced expectations and insights from PT users, often absent in closed-ended responses. This paper proposes a methodology utilizing Latent Dirichlet Allocation (LDA) to extract valuable information from open-ended survey responses, shedding light on parents' expectations regarding their children's school commute via PT. We analyzed responses from two surveys involving 448 households, focusing on parents with children in Lisbon Metro Area schools. Surveys spanned the 2017–2018 and 2018–2019 school years and pre-and post-field interventions.

The remainder of the paper is structured into five sections. Following this brief introduction, Section 2 reviews the literature on sustainable school commuting and the application of Latent Dirichlet Allocation techniques to extract information from responses to open-ended questions in mobility surveys. Section 3 outlines the methodological aspects of the research and provides an overview of the data collected through two surveys conducted in schools in the Lisbon Metro Area. Section 4 presents the obtained results that are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Literature review

2.1. Sustainable school commuting

Regarding the PT expectations, the literature indicates that this mode's choice is strongly dependent on the degree of satisfaction, especially in terms of punctuality and comfort (Papaionnou, 2017; Mouwen, 2015; Del Castillo and Benitez, 2012; dell'Olio et al., 2011; Eboli and Mazzula, 2009). Better primary school students' services are more significant in the shift to PT than for other school levels (intermediate and secondary). Such conclusions are consistent with studies reported in the literature and support an association between a child's age and the modal choice (Babey et al., 2009; McDonald, 2008a, 2008b). Also, Westman et al. (2017) demonstrated that engaging in active modes of transportation and utilizing school buses are associated with a higher satisfaction level than traveling by car. Possible explanations for this phenomenon include increased physical activity, social interactions, and the opportunity to explore the environment.

PT is part of childreńs independent mobility and essential for the five primary domains of child well-being (i.e., physical, psychological, cognitive, social, and economic), as Waygood et al. (2017) suggested. Jones et al. (2012) emphasize that public transport modes broaden young people's ability to have healthy and independent lives. Given the various potential health benefits associated with public transportation (PT), promoting its usage for adolescents' school-related travel is imperative. Possible initiatives could include rising parking prices to discourage private automobile use, improving PT infrastructure and subsidies, and changing users' perceptions of the public transport service (Mindell et al., 2021), among many other urban mobility demand management measures.

It is imperative to segment the population concurrently to effectively address the shift in mobility patterns from car-centered to sustainable commuting modes. Assuming that an approach to society must be divided into small segments to be more effective, James et al. (2017) developed a critical assessment of Individualised Marketing and Travel Blending Intervention in Canada from 1986 to 2011. According to the author, despite the efforts in all the initiatives, the positive results were insufficient to maintain political support. The initiatives need consistent success to be sustainable and potentially not target households but sectors such as companies, universities, and schools, all with specific needs.

Until a certain age, parents or caregivers in the household hold the decisive authority in choosing the mode for the school trip. Parental decision-making is perceived as a predominant factor influencing a child's commuting behavior, as posited by the hypothesized causal relationship between urban morphology and a child's school commuting (Baron and Kenny, 1986) (Bauman et al., 2002). Also, any authors corroborate the causal relationship between a childś age and the modal choice (Babey et al., 2009; Bere et al., 2008; Johansson et al., 2011; McDonald, 2008a, 2008b; Robertson-Wilson et al., 2008).

As referred before, Westman et al. (2017) showed that active modes and school buses provide higher travel satisfaction, considering the physical activity and social interactions involved. Nevertheless, raising awareness among parents and school teachers regarding these effects is essential. PT does not provide a door-to-door service. It promotes active and healthier walking trips, expanding the childhood experience by providing open spaces where young people can explore and create memorable moments (Jones et al., 2012). Likewise, enhancing environmental quality and the well-being of children presents promising opportunities to reduce car use and dependence. It can also decrease energy consumption and, more importantly, improve children's health via increased outdoor activities and by providing healthier environments (Freeman and Quigg, 2009).

Transitioning from primary to intermediate education is a strategic moment for facilitating changes in travel behavior toward more sustainable options. Moreover, interventions focusing on children living within a school distance of 2.5–3 km can promote active commuting, and maintaining this habit during secondary school contributes to contradicting the trend of decreasing physical activity

during adolescence (Cooper et al., 2012). Experimental evidence on children is restricted, but those who actively commute to and from school and subsequently adapt to an active lifestyle seem to maintain a healthy and active lifestyle thereon (Faulkner et al., 2010; Tudor-Locke et al., 2001).

Changing travel behavior towards more sustainable options is always challenging, and many projects aim to foster such changes. *The Walk to School* campaign aimed at Victoria (Australia) primary school students to commute by walking or cycling (Sahlqvist et al., 2019). Schools that participated in the campaign managed promotional materials and posters and recorded active travels during the experience. Data collected showed that these 78.628 students increased 15.5 % their active school travel. This research also concluded that the most active students were girls and students who lived within 1 and 2 km from school (Sahlqvist et al., 2019). Marketing in transport modes has also been studied using focus groups such as that in Canada, which evaluated 79 girls between the ages of 7 and 15 (Sauvage-Mar et al., 2019). The social marketing aimed to increase girls' physical activity through active modes. The study showed a segmented response: primary school students valued fun and health, intermediate school students appreciated their contribution to the environment and their opportunity to socialize, and secondary school students gave more importance to their autonomous mobility. This study is worth pointing out for the broader scope of active transport, including walking, cycling, skating, and public transport. Some other studies showed the importance of targeted campaigns for selected groups, choosing appropriate claims to their language and concerns (Lee and Kotler, 2011; Panter et al., 2008; Sauvage-Mar et al., 2019). Furthermore, other studies focused on the environmental impact of mobility (Fujii and Taniguchi, 2005; Timperio et al., 2006).

Some common obstacles to using public transport to school are distance, cost, parental chain, convenience, and level of public bus services. An alliance is required between the stakeholders (government authorities and schools) to minimize and address the challenges collaboratively (Mindell et al., 2021).

2.2. Open-ended survey questions and the use of latent Dirichlet Allocation

LDA (Latent Dirichlet Allocation) has become one of the most popular probabilistic text modeling techniques in machine learning (Wei and Croft, 2007). It has been thoroughly explained in the original paper by Blei et al. (2003), Griffiths and Steyvers (2004), Heinrich (2005), Blei and Lafferty (2009), Berry and Kogan (2010), Blei (2011), among others. Consistent with prior research conducted by Ghazizadeh et al. (2014), Mehrotra and Roberts (2018), and Roberts and Lee (2014), the initial step involved data preprocessing. This entailed importing the text from respondents' open-ended questions into a corpus data structure and subsequently removing punctuation. Words that do not contribute valuable information, such as sentence connectors (e.g., "and," "but," "the," "a," etc.), were identified as stop words and excluded from the analysis (Manning and Schutze, 1999).

Analyzing data gathered from open-ended questions offers researchers a direct view into respondents' concerns and potentially delves into their ideas on the subject more deeply (Fowler, 1995; Biehl et al.,2019; Baburajan et al., 2020). Alec et al. (2019) applied LDA models to focus group collected text data, showing that these methods can be complementary. The combination of text mining and topic modeling in LDA enables the users' suggestions to be extracted and evaluated, then used to determine the nature and extent of required changes, for example, in the PT services (Valença et al., 2023; Roque et al., 2019).

However, several topic modeling methods prove less effective when applied to short texts. Moreover, numerous challenges arise when employing topic modeling approaches for short textual data, encompassing issues like slang, data sparsity, spelling and grammatical errors, unstructured data, insufficient word co-occurrence information, and including non-meaningful and noisy words. Recently, Albalawi et al. (2020) surveyed topic modeling, which involved exploring its tools, applications, and methods. The authors tested and compared five commonly used modeling methods on short-text data to assess their effectiveness in extracting topics. Even considering short-text data issues, the results showed that LDA was one of the most effective methods for extracting meaningful topics.

This research was also designed to develop a methodology enabling the acquisition of meaningful information on school mobility based on two surveys of schoolers' parents. The survey included mainly closed-ended questions. The open-ended questions were included at the end of the questionnaire to complement previous answers, including sentences, stories, and words. Our objective was to assess the households' criticisms and recommendations for promoting sustainable transportation by improving PT services using LDA.

The decision to employ open-ended questions within the questionnaires, rather than opting for methods such as focus groups or in-depth interviews, was driven by its considerable cost-effectiveness in gathering qualitative data from a wide-ranging and heterogeneous sample. Moreover, this approach adeptly addresses several limitations associated with alternative methodologies. Notably, it mitigates challenges related to scheduling and logistical complexities. Additionally, it circumvents the inhibitions that some participants might experience in expressing personal viewpoints concerning their children, particularly within face-to-face interactions or group settings.

LDA allows to identify latent topics in the text by examining various data entities, including documents (Blei et al., 2003), images (Iwata et al., 2007), or videos (Wang et al., 2007). Topic models are probabilistic latent variable models applied to documents, leveraging correlations among words and latent semantic themes (Blei and Lafferty, 2009). It is a method for inferring topics from documents that can represent shared information among the texts collected. It is a way to obtain recurring patterns of words in textual narratives. Topic modeling uses the likelihood of words occurring in specific patterns relative to some topic by observing the frequency with which those words are used for weighing the discovered topics (Blei and Lafferty, 2009).

There is a wide range of studies using LDA to identify topics, model problems (Roque et al., 2019; Cardoso et al., 2008, among others), and solutions to a specific issue, but we were unable to identify studies using LDA to assess topics before and after field interventions and using survey open-ended questions. With the methodological approach proposed here, we aim to unfold meaningful information and support policymakers and operators to potentially incorporate new insights to improve PT services, particularly school travel behavior.

2.3. Action research and data collection

The literature lacks targeted marketing strategies to leverage school commuting by PT for younger citizens (i.e., students from primary and secondary schools), and when it occasionally exists, it is limited to single interventions and a single municipality (Cairns et al., 2004; Fujii and Taniguchi, 2005; Mitra and Buliung, 2014; Stark et al., 2019, among others). This paper is part of a broader action research project to develop a marketing mix model to leverage PT's school commuting (Queiroz, 2020b).

Our LDA modeling holds on the textual dataset collected from two surveys of 448 households whose children attend primary, middle, and high schools in three municipalities of the Lisbon Metropolitan Area (LMA). The first sought to collect data regarding the users' satisfaction, current mode choices for school commuting, and their perceptions and expectations towards public transport. The second survey sought to evaluate the impact of actions implemented in the schools involved, which aimed to stimulate students to shift to PT. The action research lasted two school years. Children and a restricted number of parents actively engaged in seven marketing events during school hours throughout the intervention program. These events were structured around the classic 4 Ps marketing mix framework, which includes Product, Price, Place, and Promotion, as introduced by McCarthy (1960). Data was collected to assess the corresponding influence on respondents' decisions to shift to public transportation before and after the interventions (Queiroz, 2020b).

It's worth noting that this study does not focus on the existing yellow school bus system, which is common in the USA but not in the EU (National School Transportation Association, 2013). There are two primary reasons for excluding this solution: firstly, despite the existence of this system in the USA, a significant number of private car journeys continue, and doubts persist regarding the cost-effectiveness of these exclusive school bus systems, as the buses often remain inactive for a considerable part of the day. Notably, European and American children with free school bus services still commute by car to school (Ewing et al., 2004). The factors influencing the ease or difficulty of adopting active modes in response to this negative trend have garnered significant attention across various research fields, including socio-demographic factors such as age, gender, income, occupation, and the environmental context, encompassing aspects like traffic and land use (Davison et al., 2007; McDonald et al., 2014; Carver et al., 2019), among others.

In contrast, the stringent schedule of school transportation, dictated by the school calendar, challenges the optimization of bus

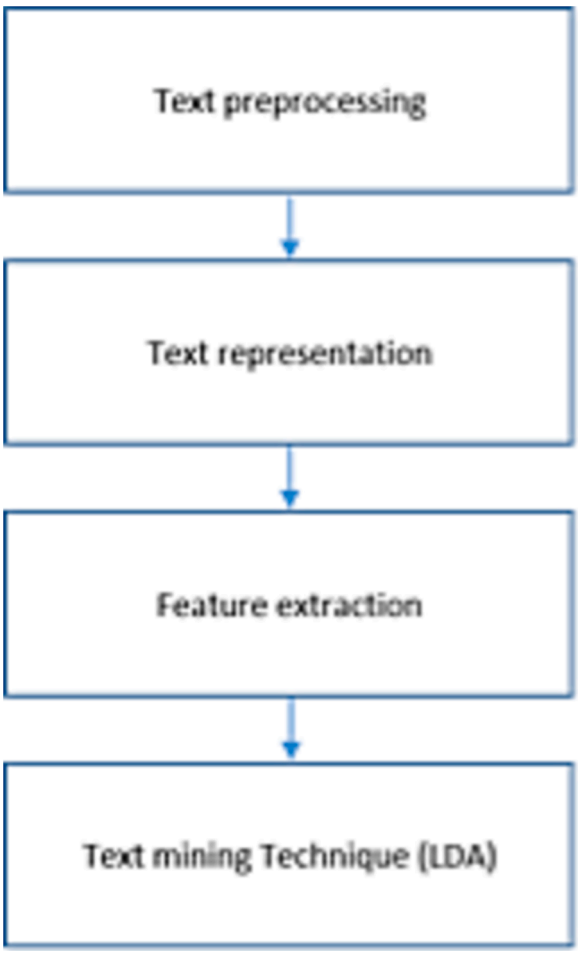

Fig. 1. Text mining and topic modeling using LDA in four key steps.

operations. Nonetheless, a plausible solution to mitigate the underutilization of standard public transportation involves strategically coordinating school transportation with other community PT services. This strategic alignment holds the potential to significantly alleviate the overall financial burden associated with this category of transportation for educational institutions (NSTA, 2013), concurrently fostering an educational agenda aimed at cultivating a greater propensity for public transportation usage among the younger population.

3. Methodology and data

3.1. Topic modeling and latent Dirichlet Allocation

LDA relies on the bag-of-words assumption (Blei et al., 2003), in which the words in a document are commutable, and their sequence is not essential, which leads to a Document-Term Matrix (DTM) that captures the frequencies of words in documents. This method relies on an unsupervised Bayesian learning algorithm, where the number of topics the model discovers is a free parameter and does not incorporate manual code into the learning procedure.

This language is structured based on latent proportions that actors may not be aware of (McFarland et al., 2013). A significant challenge in the LDA model is determining the optimal number of latent topics (K). Initially, this number is unknown but essential to initiate the model. LDA utilizes Bayesian inference to estimate the model's distribution solely based on the words within the texts. Therefore, it necessitates the pre-definition of K to commence the modeling process.

Unfortunately, no direct method exists to identify the ideal number of topics to include in the model. Consequently, various studies recommend different approaches for determining the optimal K (Zhao et al., 2015; Ravi et al., 2010; Arun et al., 2010). The most straightforward way to assess topic models is to examine the quality of each topic individually and evaluate their relevance while also considering the risk of overfitting (McFarland et al., 2013; Dyer et al., 2017).

The relationship between two words can be analyzed by tracking how frequently word X follows word Y. These two-word phrases, known as bigrams, are more informative than individual words as they provide a more comprehensive understanding of the phenomenon being studied, which, in this case, is the enhancement of the PT. It is also essential to discover the correlation between two words and determine how often words appear together and separately in the same document. The mean square contingency coefficient does this evaluation, measuring the extent and direction of correlation between two variables (Selby et al., 2014).

We used the statistical open-source software R Version 3.4.2, to perform the text mining procedure, specifically, the "tm" (Feinerer et al., 2008) and "topic model" packages (Grun and Hornik, 2011). The former provides text mining functions, while the latter implements the LDA algorithm. The text's acronyms "pt" and "PT" always refer to Public Transport. In Fig. 1, we seek to stylize the steps taken with this method (LDA).

The following section provides an overview of the flowchart for text mining and topic modeling, detailing the following steps:

- *Step 1 - Text preprocessing*: In this phase, we cleaned and prepared the text data by executing essential tasks like tokenization, converting text to lowercase, eliminating stopwords, and applying stemming or lemmatization.
- *Step 2- Text representation*: After preprocessing, we represented the text using techniques like Bag of Words (BoW). Word clouds were created from the terms with the highest frequencies in the processed text.
- *Step 3- Feature extraction*: The technique of using bi-grams was applied in this step to capture more meaningful relationships between words and provide additional context compared to individual words. They helped in understanding the structure and semantics of a piece of text.
- *Step 4- Text mining Technique (LDA)*: The Latent Dirichlet Allocation (LDA) technique within text mining was employed to unveil latent topics in the databases. The primary objective was to identify concealed thematic structures and reveal the distribution of topics across the entire corpus.

3.2. Surveys

Open-ended questions in surveys may be used to explore deeper information and expectations in Public Transport (PT). We developed two surveys to collect information about school commuting before and after the fieldwork. The parents and children were informed of the purpose of the surveys by attaching explanatory letters. The first survey aimed to assess the level of satisfaction with PT and identify the improvements needed. The second survey aimed to determine the reactions to the interventions and their effectiveness in the shift to PT when commuting to school. The fieldwork consisted of implementing *marketing experiments (that we refer to as "stimuli"* for a behavioral change) associated with the 4 Ps of Marketing and collecting information to evaluate the corresponding impact on respondents' decision to shift to PT. The marketing events included public debates, bus papers (BP; Learning Public Transport Rally), stands (which are Product-related), free passes (Price-related), Traffic Snake Game, think tanks (Place-related), and the use of a transport routing App (Promotion-related).

Bus Papers (BP) are a competitive game that requires students to visit different locations and perform some activity to move to the next successfully. In BP games, students must walk, run, take the bus, a train, or other PT modes from one location to the next. Activities can be answering questions related to PT, reading printed materials, using storytelling techniques, singing a popular song, dancing, or other activities. They are then awarded points at each location depending on the quality of the answers, and the winners receive a prize (e.g., a free PT pass). The BP combined general culture, essential urban mobility culture, speed, humor, and other skills. It is mainly outdoor. The aim is to expose children to new PT experiences and potentially improve the PT operatorś image and

communication.

These two surveys collected empirical data for this study on ten elementary, middle, and high schools (6–18 years) of three municipalities of the LMA. Then LMA is home to almost 3 million inhabitants, which accounts for nearly 30 % of the Portuguese population, all concentrated in just 3.3 % of the national territory (INE, 2021). It stands as the most significant economic agglomeration in Portugal. The private car is the primary mode of transport, with 56.3 % of weekday trips, while public transport accounts for only 15.8 % of those trips (IMob, 2017). Importantly, LMA's students are eligible for a complimentary public transportation pass if their residence exceeds 3 km from the school. This provision applies specifically when students attend schools outside their residential catchment area, either due to unavailability of space in the neighboring school or because the desired learning topic is not offered there. Additionally, this criterion extends to students whose academic level cannot be accommodated by the available parish schools.

Public and private schools were included in the case study, and we also aimed to discuss how different school environments might influence the type of travel of families and students. The surveys were paper-based, following the school's board of direction advice, as it would be more effective based on their past experiences. The school's boards of directors and teachers distributed the questionnaires to children and households. The Portuguese National Data Protection Commission approved the surveys' procedures and design (CNPD Ref. 02.02-Ofic. n° 5935/2018).

The first Survey (February 2018) included the following sections: socio-demographic information (parents and children); mobility routines (parents and children); public transport assessment (parents only); choices between car and public transport (parents only); personality type (parents only); and environmental awareness and attitudes (parents only). The questionnaire also included an open-ended question at the end, where respondents (parents or children or parents on behalf) were asked: "If you wish to leave a comment, please use the text box below?".

A second survey was realized in May 2019 and consisted of four main sections: socio-demographic information (families), public transport assessment (parents only), shift to PT assessment (parents and children), when and why, and evaluation of the measures implemented (parents and children). Again, an open-ended question was used to finalize the questionnaire, where respondents were asked the same question. These two surveys were cross-sectional, and because of the need to respect the regulations protecting personal data, it was impossible to establish any connection between the first and second surveys.

The overall response rate was 41 % (1640) and 43 % (1760) for the first and second surveys. After the data cleaning, the collected responses included 448 valid open-ended answers, 103 (23 %) responses from the first survey, and 345 (77 %) from the second. The questionnaires' different lengths can explain the difference in response rates between the two surveys: ten pages in the first survey, while the second had four pages. The original texts from responses were written in Portuguese and translated to English by a single researcher using a single online translator to ensure consistency in the overall process. The survey responses were aggregated in a single corpus since the questions were identical. The open-ended questions were not mandatory, and 448 valid responses were collected from 3400 participants (13 %).

3.3. Data and descriptive statistics

Table 1 presents the respondents' socio-demographic descriptive statistics who answered both open-ended survey questions. The responses collected in both surveys did not result in a skewed demographic profile. Parents correspond to 90 % of the respondents, while the remaining are other caregivers (e.g., grandparents). 46 % of the survey respondents were 35 to 44 years old, while 39 % ranged between 45 and 54. Some respondents are younger and fall into the age group of under 24, the parents of primary school students. 68.5 % of the respondents are women. The majority of respondents (70.9 %) had full-time jobs. Concerning education, 40 % hold a graduation degree, while 38 % attend high school. Of the total number of surveys, 18 %, 42 %, and 40 % were from primary (6–10 years), intermediate (11–12 years), and secondary (13–18 years) schools, respectively.

Consistent with prior research conducted by Ghazizadeh et al. (2014), Mehrotra and Roberts (2018), and Roberts and Lee (2014), the initial step involved data pre-processing. The topic analysis focused on the spreadsheet containing the reviews of the respondents. We reorganized some data information to create a document-term matrix (DTM) that can be processed via topic modeling, and we made some assumptions for pre-processing choices. The DTM is an input to the LDA topic modeling to get the most relevant topics (Blei et al., 2003).

The process of text pre-processing in this study included tokenization, converting words to lowercase, removing punctuation, and removing stop words. This entailed importing the text from respondents' open-ended questions into a corpus data structure and removing punctuation. Words that do not contribute valuable information, such as sentence connectors (e.g., "and," "but," "the," "a," etc.), were identified as stop words and excluded from the analysis (Manning and Schutze, 1999). In instances where words are presented as "bus" "417″ and "bus" "408," we performed concatenation as follows: "bus417″ and "bus408." This concatenation was implemented to prevent word separation, ensuring clarity in the intended meaning explicitly denoting the reference to respective buses. Table 2 briefly characterizes the corpus of the two databases (Survey 1 and Survey 2).

4. Results and discussion

4.1. Relationship between words

Studies have shown that a word cloud is a simple but effective illustration of data (Gao et al., 2016; Sun and Yin, 2017). Fig. 2 shows the word cloud of reviews from respondents. A word cloud determines the most frequent terms in a corpus. Let $p_{x,y}$ be the rate at which the word x occurs in document y, and p_y be the average rate across n documents ($\sum_y p_{x,y}/n$). When comparing clouds, each word's size

Table 1
Summary socio-demographic descriptive statistics of the open-ended questions' respondents.

Variable	Description	Overall *Freq. (%)*	Survey 1 *Freq. (%)*	Survey 2 *Freq. (%)*
Parents	yes	90	87	90
	no	10	13	10
Age	$\leq$ 20 years	1	0	2
	20–24	1	1	0
	25–34	6	5	6
	35–44	46	41	48
	45–54	39	43	38
	55–64	6	10	6
	$\geq$ 65	1	0	0
Female	yes	67	67	68
	no	33	33	32
Employment	yes	82	18	87
	no	18	17	13
Level of education	Primary	22	23	22
	Secondary	38	39	37
	Grade level	40	38	41
Income	Live without financial restrictions	30	26	33
	Live modestly	56	60	53
	Live with financial restrictions	14	14	14
	0	10	15	11
Number of cars	1	40	35	40
	2	42	44	40
	3	5	2	5
	$>$ 3	3	4	4
	Primary school	18	20	18
School level	Intermediate school	42	20	50
	Secondary school	40	60	32
	Cascais	22	28	20
Municipality	Oeiras	61	59	62
	Sintra	17	13	18

Table 2
Corpus characterisation (Survey 1 and Survey 2).

	First survey	Second survey
# respondents	105	343
Average of words per sentence (Standard deviation)	3.73 (1.68)	4.39 (2.29)
Maximum number of words per sentence	9	12

is mapped to its maximum deviation $max_x\left(p_{x,y} - p_x\right)$ (Subasish et al., 2010; Das et al., 2016). The more frequently words are used, the bigger they appear in the word cloud. World cloud is a visual method of obtaining tangible insights from qualitative data. In our study, the word cloud clearly shows five prominent themes related to public transport school commuting: school, safety, comfort, increase, and school bus schedules.

Fig. 2 shows the most frequent terms of the first survey (on the left side), corresponding to the words "bus", "school", "safety", "stop", "punctuality", and "schedules". These are generic concerns with the performance of many PT systems. It also shows the most

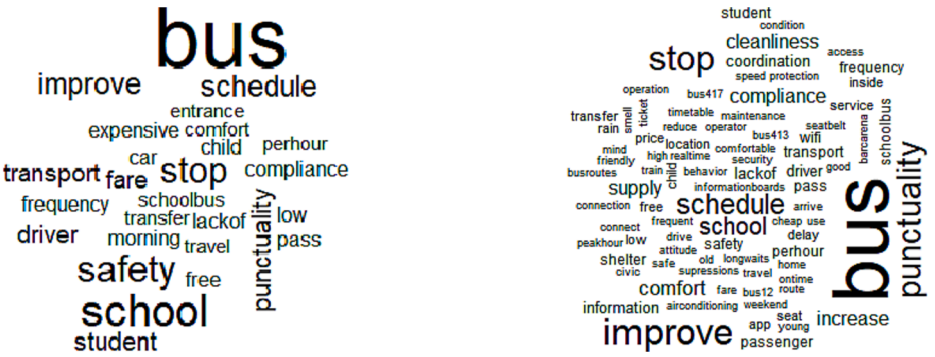

Fig. 2. Word cloud of reviews of the first survey (left) database and second survey database (right).

frequent terms of the second survey (on the right side), corresponding to the words: “improve”, “punctuality”, “schedule”, “supply”, “compliance”, “cleanliness”, “comfort”, “drivers”, and “information”. The figure unveils new items: “app” (a friendly public transport route planner app suitable for young people), “access”, “seatbelts”, “seats”, “busroutes”, and “transfers”. The analyzed text shows emerging concerns regarding PT, such as the need for real-time information and comfort inside and outside vehicles.

Fig. 3 plots the combination of related words for the recordset[1] of the first survey. The relationships are directional (ticked with an arrow). There are some words, such as “pt”, “bus” and “school” that form common centers of nodes. We also see pairs and triplets that include common phrases related to PT concerns (e.g., “low fare”). On the other, some solutions are also recommended – for instance, “compliance schedules”, “coordination school bus schedules”, “student discount”, and “ free pass.”.

The word “safety” is followed by “convenience” and preceded by “comfort”. Characteristics and concerns associated with PT emerge in this sequence “comfort, safety, and convenience”. The word “school” is followed by “entrance” and preceded by “children”, which reveals the concern with coordinating the timetable of the buses to ensure that students arrive in class on time.

We also see pairs with messages related to prices (lower fares and free passes). It also shows that the messages are linked to schedules (compliance with schedules and coordination between schools and bus schedules).

Fig. 4 depicts the words most correlated with “bus”, “bus_stop”, “improve”, and “schedules” which are present *in opinions and comments* recorded. The words most correlated with “bus” were “few” and “perhour”, and associated with the “bus-stop”. The words most correlated with the word “improve” relate to problems with primary attributes of the PT, such as “punctuality,” “bussafety” and “attractive”, hence covering topics of reliability and aesthetics. The “bus_stop” element exhibits a strong correlation with various terms related to the absence of ticket-selling machines for transport tickets, the proficiency of drivers upon arrival at stops, and the location of bus stops, particularly in proximity to the specified school “Castelo Branco” in the Oeiras municipality. Additionally, examining words correlated with the term “schedules” reveals issues about compliance, managerial accountability, coordination of train schedules with bus services, and an overall inadequacy to meet the demand. Despite its apparent status as a stopword, the inclusion of the term ’only’ is intentional, serving to underscore and fortify patterns within substantive content-bearing words. In this context, it accentuates the constraint of limited frequency, explicitly highlighting the occurrence of buses solely once per hour.

Fig. 5 depicts the network of bigrams, corresponding to relationships between words in our second survey dataset. Compared to the first survey, this large diagram highlights that the number of nodes is substantially more significant, increasing word relationships and expanding the interdependence between the whole lexical system used. Also, the dataset of the second survey was larger than that of the first survey. We identified the following centers of nodes “improve”, “punctuality”, “bus” and “bus_stop”.

The node “bus-stop” centralizes issues such as information, app, shelters, and WIFI location tracking. On the other hand, the nodes “bus”, “improve” and “punctuality” are interlinked. The diagramalso shows that the messages are linked to schedules (compliance with schedules and coordination between schools and bus schedules). The PT supply appears described in three ways: lack of pt, increase supply, and bus per hour. Free passes also seem to be necessary. The issues of free passes and increasing the supply are mentioned. Another aspect discussed alone, as if the only solution, was “schoolbuses”.

In the second survey, various solutions to transport problems were mentioned:

- Operational management of the transport service: frequency, punctuality, routes, location of stops, schoolbuses;
- Communication: app, teleinformation panels;
- Comfort: wifi, shelters, cleanliness; care of bus interiors;
- Liaison between the school and transport operators: school and bus timetables.

In Fig. 6, the words most correlated with “bus” were “more”, “perhour”, “inside” and “avoid”. The words associated with the “bus-stop” were “shelters”, “rain”, “location” and “app”. The words correlated with the word “improve” relate to problems with primary attributes of the PT, such as “punctuality,” “cleanliness”, “smell” and “comfort”, hence covering topics of reliability and comfort.

The “bus_stop” element is highly correlated with words regarding aspects of comfort (“shelters” and “rain”) and optimization by informing the PT location in real-time through an app (“location” and “app”). Also, examining the words correlated with the word “punctuality”, it is possible to identify issues regarding improvement and other needed characteristics of PT, including cleanliness, comfort, and bus safety.

4.2. Topic models and interpretation

The first decision when using LDA is to identify the number of topics or define the K parameter. We use the R package “ldatuning” (Nikita, 2016) to choose the most adequate K value, i.e., the number of topics. Increasing the number of topics producing detailed partitions can result in a less suitable model because it becomes impossible for humans to differentiate between numerous topics (Chang et al., 2009). As such, the increase in fit is sometimes due to overfitting (Dyer et al., 2017). Ultimately, the choice of models must be determined by the researched questions, i.e., the privileged interpretability of the topics instead of statistical improvements. DiMaggio et al. (2013) suggest that the process should be empirically controlled, in that if the data are unsuitable for answering the analysts’ questions, no topic model will produce a valuable data extraction.

No precise or correct number of topics exists (Roque et al., 2019). Figs. 7 and 9 illustrate the impact of the number of topics K on

[1] Consisting of a group of database records from a base table resulting from a query in the survey.

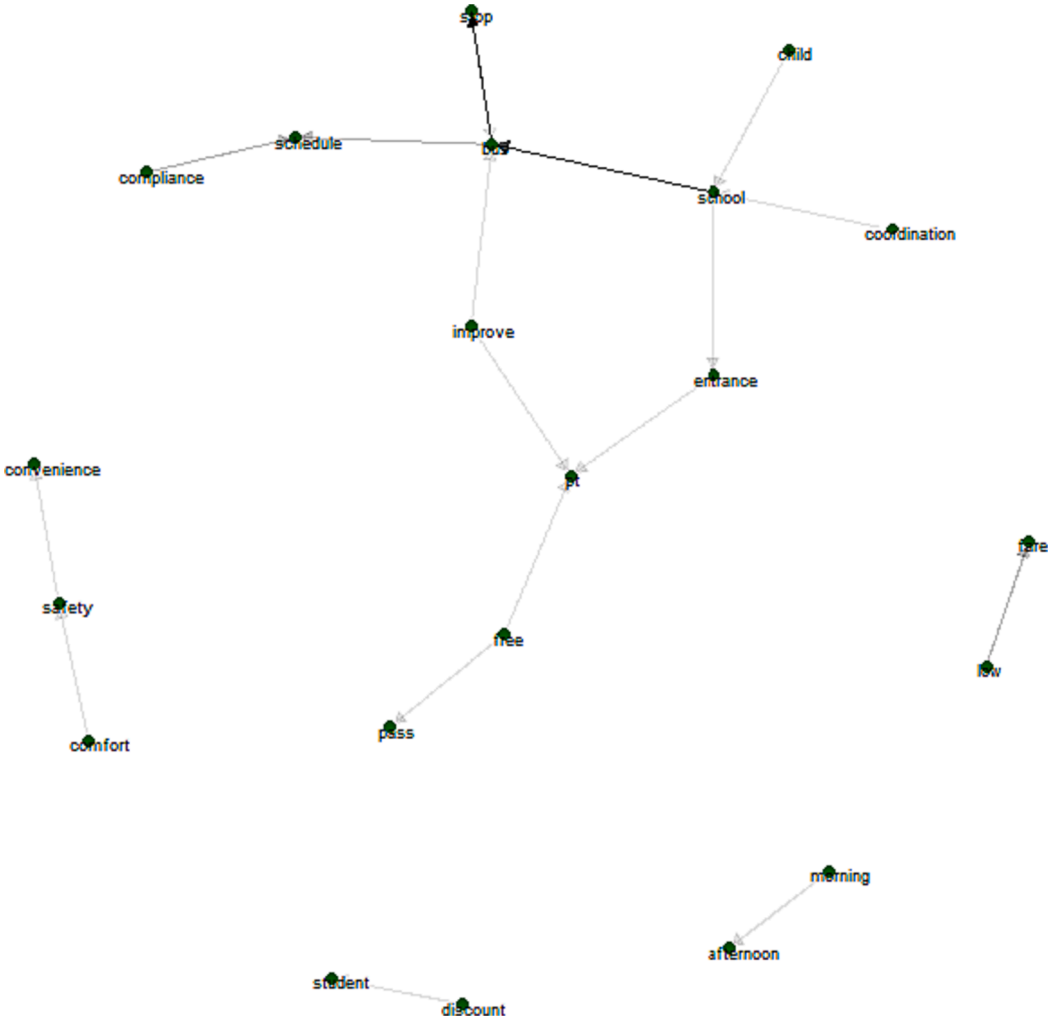

Fig. 3. Directed graph of common bigrams in the database of the first survey.

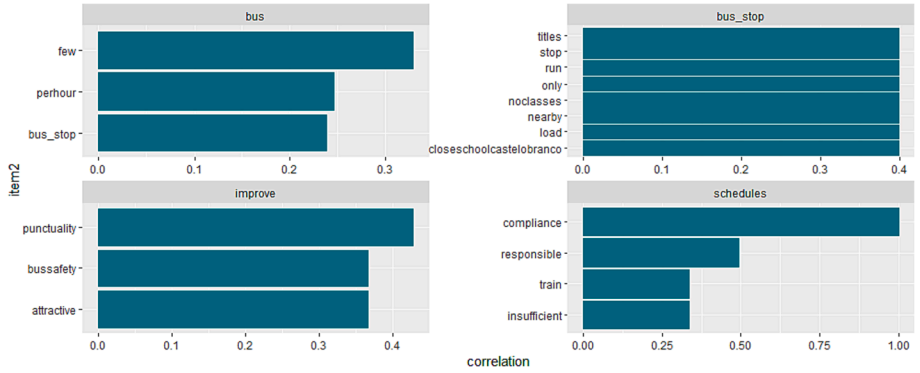

Fig. 4. Words associated with "bus," "bus_stop," "improve" and "schedules" in the opinions and comments recorded in the first survey.

each LDA model, where K ranges from two to 25. The expectation–maximization method limits the number of topics to 25. LDA's latent variables estimation uses the Gibbs sampling algorithm since it overcomes the problem of obtaining samples from complex probability distributions using random numbers (Mackay, 2005). The sampling is done successively until the tested values approximate the target distribution (Steyvers and Griffiths, 2007).

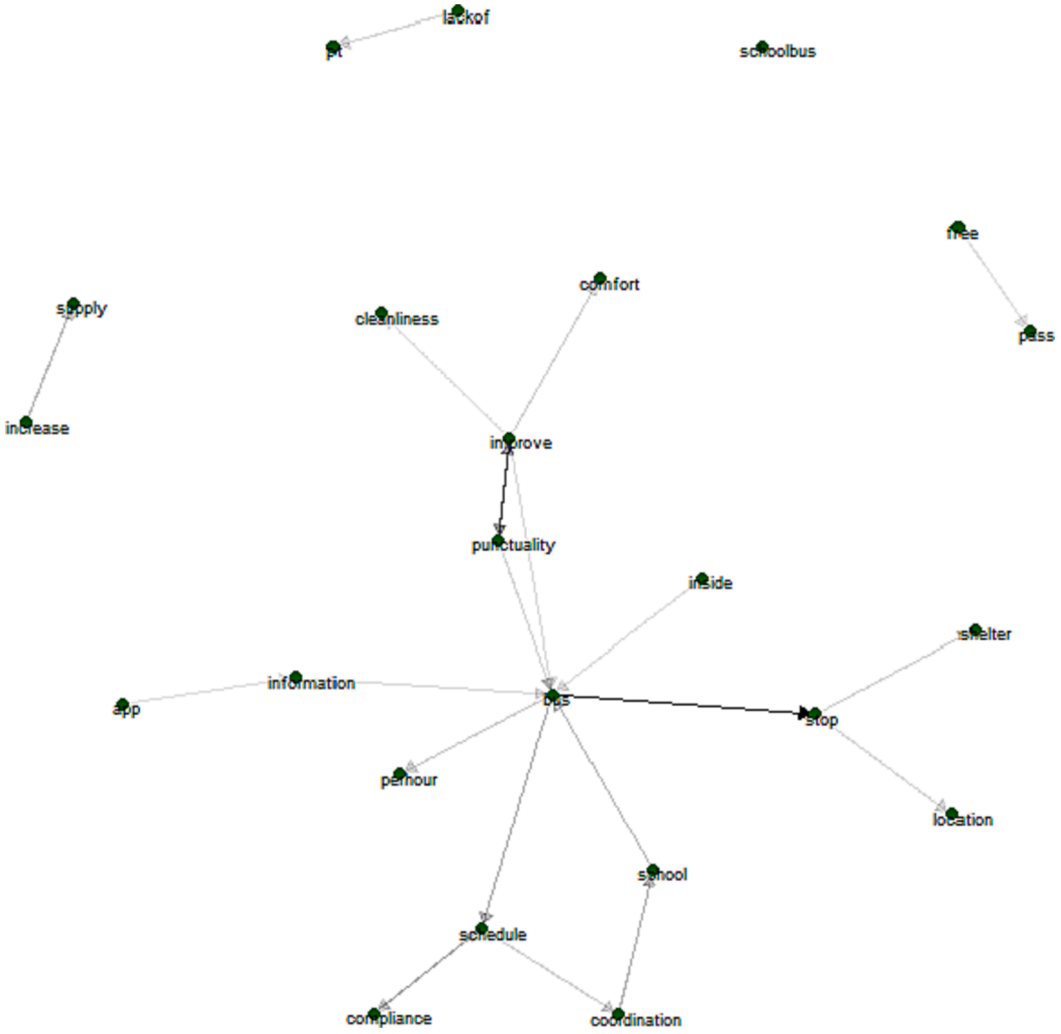

Fig. 5. Directed graph of common bigrams in the recordset of the second survey.

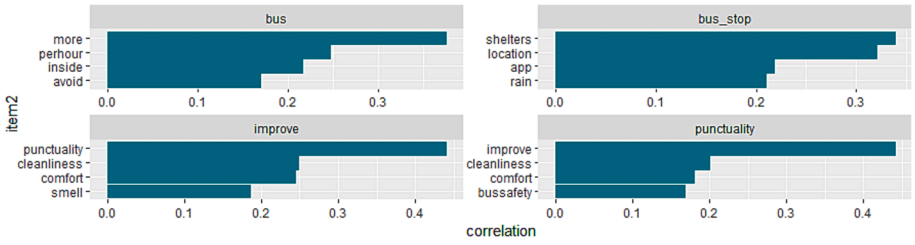

Fig. 6. Words associated with "bus," "bus_stop," "improve," and "punctuality" are in the opinions and comments recorded in the second survey.

According to McFarland et al. (2013), a simple way to evaluate topic models is to analyze the appropriateness of each topic. A higher number of topics may result in smaller and potentially meaningless dimensions. Fig. 7 illustrates the expectation–maximization results obtained for different K topics for the first survey. We conducted several model runs, ranging from 6 to 13 topics. We determined that the model with six topics exhibited the most efficient clustering of the first survey dataset, as indicated by the word distribution per topic.

Fig. 8 presents the probabilities of topic-specific words (β) for the six topics within the record set for a more in-depth comprehension of LDA's latent topics. By comparing the top 5 words from each topic's word distribution and their prevalence across topics, the word "bus" is more likely to be related to Topic 2 (20 %) than Topic 4 (8 %). Also, "safety" has a 15 % likelihood of being connected to Topic 5. Through a comprehensive analysis of these topics, we discerned a shared thread, i.e., a broad concern for preconceptions related to

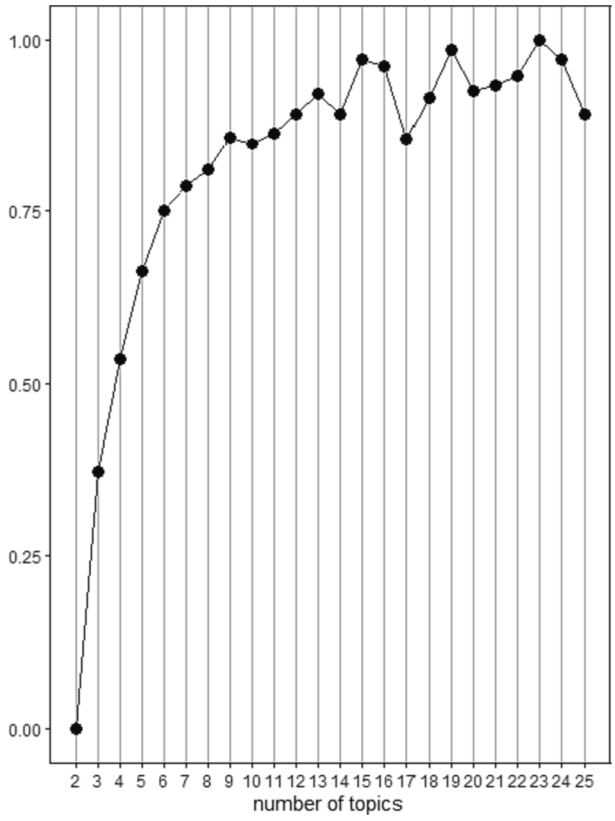

Fig. 7. Determining the number of latent topics (K) for the opinions and comments record set in the first survey.

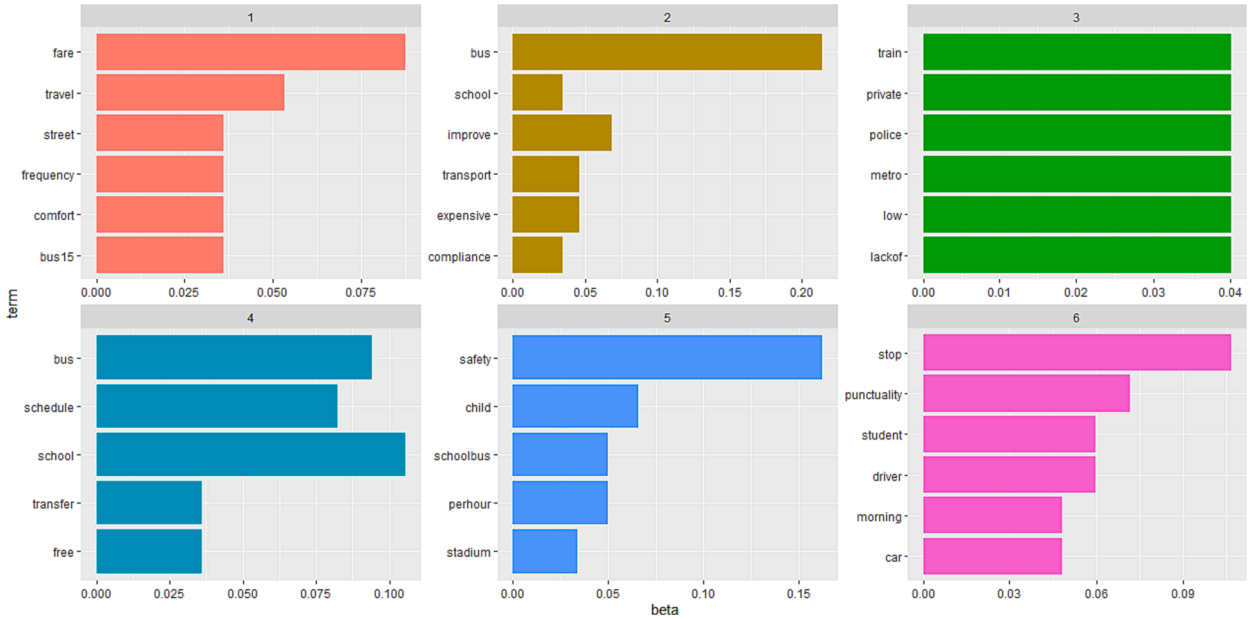

Fig. 8. Topic-specific word probabilities for the opinions and comments in the recordset of the first survey.

public transportation. These preconceptions include aspects such as security and service performance.

Similarly to Fig. 7, Fig. 9 illustrates the expectation–maximization results obtained for the second survey's different up to 25 topics. Again, the LDA models stabilized after reaching six topics, leading to a more consistent set of topics from an interpretation perspective.

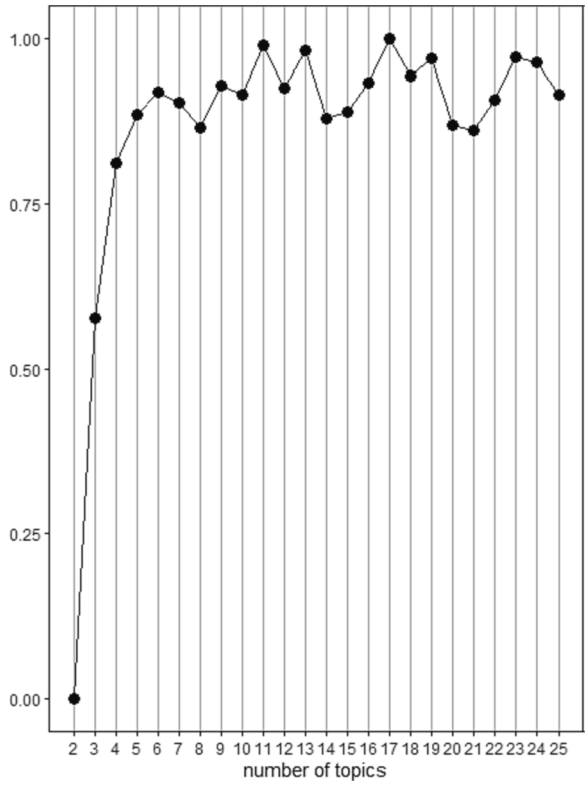

Fig. 9. Determining the number of latent topics (K) for the opinions and comments record set in the second survey.

In Fig.10 and like in the first survey, "bus" is likely related to the topic 1. In this second survey, bus scheduling is likely related to several topics. The word "schedule" ranks first in Topics 3, 4, and 6 (8 %, 16 % and 8 %, respectively). Topic 6 also includes the word "timetable" suggesting that parents would expect a better alignment of decisions between schools and operators to improve the quality of PT services for the children. Other concerns also arise related to the performance of the PT services (frequency, cleanliness, location of stops, drivers' behavior).

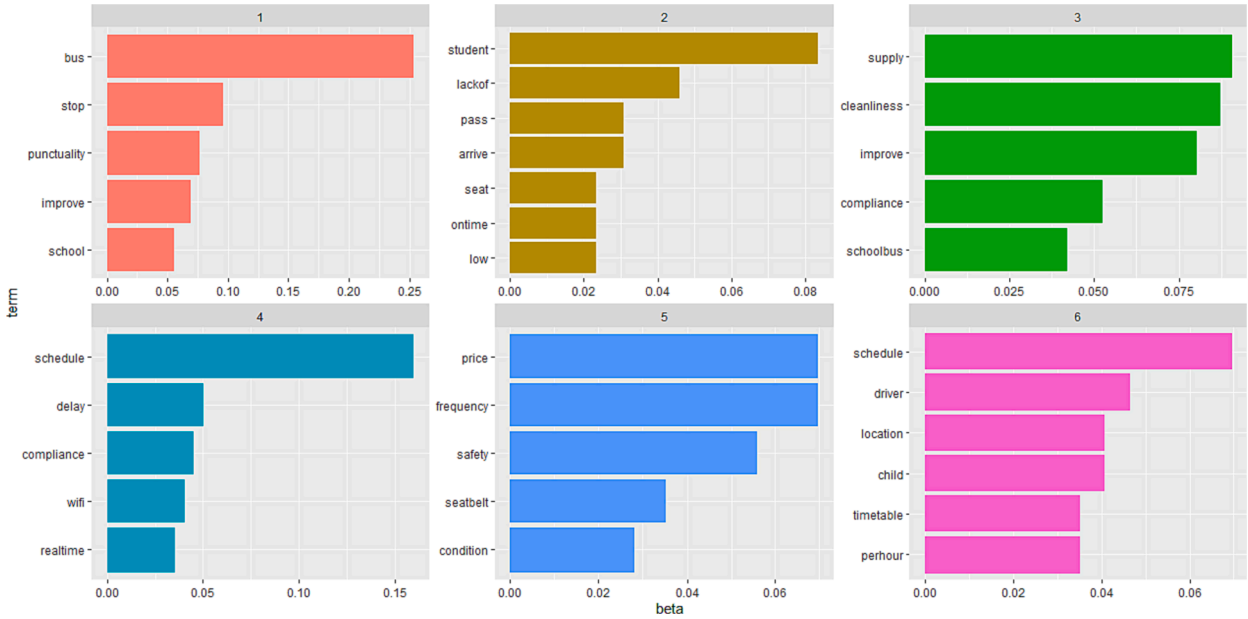

Fig. 10. Topic-specific word probabilities for the opinions and comments in the recordset of the second survey.

Table 3 shows the six extracted latent topics for the opinions and comments on PT based on survey 1. Each topic contains all words in the corpus, though with different probabilities. The extracted topics capture a meaningful data structure consistent with the class designations provided. This analysis enabled the identification of specific words to be used as topic labels. The labels chosen intend to give the best interpretability using the semantic similarity measure between the top five words of each topic's distribution.

Regarding the second survey, the five words that are more likely related to the topic are listed in Table 4. The topics were labeled based on the selected words for each topic.

Table 3 and Table 4 topics shared common concerns captured from both surveys. These include the need for service improvements, issues related to service quality such as frequency and punctuality, elevated transport costs, and the existence of school buses. The topics of the second survey suggest that respondents shed light on possible solutions to enhance the transport service. These encompass aspects such as cleanliness, installation of seat belts, improved comfort, increased seating capacity, improving bus stops and their location, integration of WiFi and a dedicated app, and better coordination between school schedules and PT timetables while enhancing common PT factors (e.g., price and frequency).

To substantiate these suggestions, we have transcribed some testimonials from the responding parents: "PT doesn't obey timetables, and stops aren't close to primary schools"; "More buses to prevent overcrowding, better operation, they shouldn't close the doors, and they should wait; drivers drive too fast"; "The location of public transport stops doesn't serve all passengers"; "Improve bus punctuality"; "Enhance comfort, safety, and timetables.".

As mentioned in Section 3.2, we conducted field experiments between the two surveys. These experiments consisted of marketing actions aimed at increasing experiences with public transport. Still, other actions were not implemented. For instance, operators should have improved the PT services to meet the expectations of the children's parents and caregivers. For example, service frequency should increase, bus routes should be redesigned to serve primary schools better, and school schedules and bus timetables should be better coordinated, among other requirements stated by the respondents of the second survey.

The combination of text mining and topic modeling enables the extraction of opinions and recommendations regarding the users' perception of PT performance that close-ended questions might not capture. These users' statements complement and underline the necessary and perhaps more prominent improvements required to potentially increase the modal share of PT.

This study provides evidence that the LDA is appropriate for identifying those opinions and recommendations. With this method, we can streamline the analysis of the open-ended questions regarding the interpretability and context of the words used and capture the traveler's stated concerns regarding PT use. Thus, including open-ended questions in travel surveys provides the researchers with additional evidence of the respondents' perspectives that complement standard closed-ended questions. In parallel, when we collect responses, we can automatically view trends or spot information that stands out with word clouds and graphs. This technique complements other data analyses, such as exploratory and confirmatory ones.

Our results support the theory that PT expectations make the difference in PT choice when commuting (Papaionnou, 2017; Mouwen, 2015; Del Castillo and Benitez, 2012; dell'Olio et al., 2011; Eboli and Mazzula, 2009). We concluded that, after being exposed to marketing interventions, respondents' statements shift from criticism and opinions regarding PT performance to proposing specific improvements to PT services. Our conclusions resonate with dell'Ollio et al. (2011) who concluded that the smaller the gap between the expected and the experienced service, the higher the satisfaction of PT users. Another study by Ramos et al. (2019) yielded similar results, indicating that a closer match to anticipated service tends to result in greater user satisfaction. Moreover, our results suggest t PT operation attributes (for instance, frequency and schedules) must be improved, corroborating Queiroz et al., (2019; 2020b). Also, Laura and Gabriela (2009) assessed the bus services quality, evaluating the significance and satisfaction of 26 attributes. Frequency and timetables emerged as particularly crucial, surpassing the importance of other factors. Mouwen (2015) further reinforces the importance of punctuality and frequency in assessing bus service quality. Also, Luigi et al. (2011) research emphasized the substantial weight assigned to waiting time. While various studies discuss improvements in transportation, the fundamental characteristics of the service—frequency and punctuality—consistently stand out as crucial attributes in all these investigations.

Interestingly, we stress that the respondents explicitly reinforced their opinions and comments regarding PT services despite having already responded to the "Satisfaction with PT" evaluation scale. On a scale of 1 to 7 Likert Scale (where 7 is the best rate), the overall average satisfaction was 3.1 (i.e., below the median value of the scale). PT stop proximity (average 3.4), and payment methods (average 3.3) were the variables with higher satisfaction (still below median lasso). In contrast, ticket/pass cost (average 2.6) and frequency (average 3.0) had the lowest satisfaction ratings. Unexpectedly, respondents did not refer to any necessary improvement in bus flexibility and tracking, as found in a previous study by Queiroz et al., (2020a) for the same case study and population.

These findings have broadened the respondents' concerns as they did not limit themselves to answering closed survey questions. They wanted to express opinions and recommendations regarding PT and their mobility expectations in their language. We analyzed the two databases separately before and after the field interventions. We identified different topics, suggesting that the field interventions influenced how the respondents reckoned on the PT service problems.

Regardless of some topics appearing in both surveys, the marketing initiatives affected the respondents' perceptions as we could observe a shift of vaguer concerns in the first survey compared to more precise and targeted comments in the second survey. The emerging topics raise issues that need to be resolved to leverage PT. One plausible strategy involves practically implementing enhancement packages corresponding to each subject. For instance, Topic 1 (Survey 2) focuses on enhancing PT infrastructures and services, where improvements should aim at stops, punctuality, and timetable adjustments. Concerning Topic 2 (Survey 2), reexamining students' pass adequacy is necessary, along with adjustments of timetables with school schedules and improving students' quality of life (e.g., arriving at school on time). Additionally, efforts should be made to provide buses with higher capacity and comfort (for instance, more seats). Implementing packages of measures aligned with the themes of the topics would facilitate tracking suggestions and defining their scope.

Table 3
Extracted Latent Topics (first survey) with keywords (opinions and comments recordset).

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Transport fares	*Improve school transport*	*PT dissatisfaction*	*Schedules & transfers*	*Schoolbus for safety*	*Morning commute & PT efficiency*
fare	bus	train	bus	safety	stop
travel	school	private	schedule	child	punctuality
street	improve	police	school	schoolbus	student
frequency	transport	metro	transfer	per hour	driver
comfort	expensive	low	free	stadium	morning
Bus15	compliance	lack of			car

Table 4
Extracted Latent Topics (second survey) with keywords (opinions and comments recordset).

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
PT services performance	*Students' PT expectations*	*PT Quality*	*PT services and connectivity*	*PT travel factors*	*Schoolbus logistics*
bus	student	supply	schedule	price	schedule
stop	lack of	cleanliness	delay	frequency	driver
punctuality	pass	improve	compliance	safety	location
improve	arrive	compliance	wifi	seatbelt	child
school	seat	school bus	realtime	condition	timetable
	on time				per hour
	low				

5. Conclusions

Our research aimed to contribute with an effective analytical method to the research field of PT, particularly in school commuting. Capturing potentially resourceful information from open-ended responses (often disregarded) can have multiple applications with immediate usefulness to PT managers, marketers, and policymakers. These can be quick wins since they are resourced directly from the users' stated concerns.

Policy implications can be derived from this work regarding the promotion of more sustainable school commuting using PT. Increasing stakeholder involvement can be achieved by raising awareness and subsequently translating this awareness into decision-making, particularly by addressing users' expectations and problems related to PT services. Collecting parents' perceptions is attractive to policymakers, as it allows for exploring contrasting opinions and a rich blend of viewpoints from this segment of stakeholders. Open-ended questions prove particularly relevant in this context. Moreover, it effectively captures the complexity of policy implementation contexts and processes, making it suitable for ex-ante and ex-post evaluation approaches, i.e., the before and after perceptions of actions and interventions.

This study holds significance as it has provided a platform for the parents of school students to voice their perspectives, enabling us to discern the solutions they propose to address specific transportation issues. Some interventions could be incorporated into a school commuting policy (e.g., coordinating transportation schedules with the school's timetables). Secondly, PT experiences could be further explored within schools. Also, it is crucial to bring stakeholders together and embrace the challenge of a more sustainable school commuting for the students. Finally, a collaborative approach would guarantee that the concerns of all involved parties are considered, facilitating the formulation of comprehensive solutions.

We confirmed the added value of text mining and topic modeling in their ability to efficiently and objectively extract meaningful insights from reviews, providing a deeper understanding of the respondents' perceptions and preferences than traditional methods. Morevover, we explored the following positive aspects of the LDA methodology. It allows for an effective comparison of open-ended responses of surveys collected before and after marketing initiatives, which can potentially capture opinions and perceptions of the PT service performance, complementarily to close-ended questions. By doing so, we could detect changes in the stated concerns, where respondents gave more detailed pieces of information and pointed out the way the operators should focus their service improvements. Also, we could identify the main topics and words and transform them into valuable insights through latent topics.

One practical use of this research could be to give the topic prevalence to the operators to improve operations. Future applications might include using the word topics emerging with the method proposed here in future complementary closed questions in the follow-up surveys to monitor the performance improvements of PT services. Another quick-win application would be to provide the marketers with text mining and topic modeling results to get deeper and more complete insights into the PT users' perceptions, opinions, and recommendations. For instance, our case study highlighted that PT user's increased use could be stimulated by improving the bus stops (location and shelter comfort), the driving quality, online and real-time access to transport applications, the bus comfort, punctuality, and intermodality of the supply, enabling fare pricing and coordination of schedules between school and buses.

The method used here can also be applied to other text sources, including social-media communications such as Twitter feeds of interview records, and it can improve the research field by broadening the traditional information sources. Implementing LDA in large-scale studies promoting sustainable transportation can significantly enhance the efficiency and cost-effectiveness of text analysis,

thereby saving valuable time and monetary resources. For future work, we intend to include secondary data sources in the LDA analysis (e.g., industry reports, other relevant surveys, and school reports) to compare with the results from the survey data. Also, we intend to explore different solutions to cope with short-text data issues by creating more extensive pseudo-document representations from the original documents. It would also be interesting to apply models that link topics to specific issues or avoid prematurely determining the number of topics. Traditional unsupervised topic models like LDA are limited to utilizing only the discrete bag-of-words representation. They cannot take advantage of any metadata accessible for each questionnaire. Thus, other modeling approaches, such as structural topic models (STM), can be explored. STM allows the researchers to discover themes from documents and estimate how the topic relates to the document metadata.

CRediT authorship contribution statement

Mariza Motta Queiroz: . Carlos Roque: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Filipe Moura:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Joao Maroco:** Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to the three anonymous reviewers for their valuable contributions.

*Disclosure **statement***

The author(s) declared no potential conflicts of interest concerning this article's research, authorship, and/or publication.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This research is part of the activity at the Civil Engineering Research and Innovation for Sustainability (CERIS) research center. FCT funded the work in the framework of the following project: UIDB/04625/2020.

Authors' contributions

The authors confirm their contribution to the paper as follows: *Study conception and design*: Mariza Queiroz, Carlos Roque, Filipe Moura; *Data collection*: Mariza Queiroz; *Analysis and interpretation of results*: Mariza Queiroz, Carlos Roque, Filipe Moura, João Marôco; *Draft manuscript preparation*: Mariza Queiroz, Carlos Roque, Filipe Moura, João Marôco. All authors reviewed the results and approved the final version of the manuscript.

References

Albalawi, R., Yeap, T.H., Benyoucef, M., 2020. Using topic modeling methods for short-text data: a comparative analysis. Frontiers in Artificial Intelligence 3, 1–14.
Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N. (2010) On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. Proceedings of Advances in Knowledge Discovery and Data Mining, 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Lecture Notes in Computer Science 6118.
Babey, S.H., Hastert, T.A., Huang, W., Brown, E.R., 2009. Sociodemographic, family, and environmental factors associated with active commuting to school among US adolescents. J. Public Health Policy 30, S203–S220.
Baburajan, V., Silva, A.J., Pereira, F.C., 2020. Open-Ended Versus Closed-Ended Responses: A Comparison Study Using Topic Modeling and Factor Analysis. IEEE Transactions on Intelligent Transportation. https://doi.org/10.1109/TITS.2020.3040904.
Baron, R.M., Kenny, D.A., 1986. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. J. Pers. Soc. Psychol. 51 (6), 1173–1182.
Basington, H., 2008. Travel Socialization: A Social Theory of Travel Mode Behavior International. Journal of Sustainable Transportation 2 (2), 91–114.
Bauman, A.E., Sallis, J.F., Dzewaltowski, D.A., Owen, N., 2002. Toward a Better Understanding of the Influences on Physical Activity. Am. J. Prev. Med. 23 (2).
Bere, E., van der Horst, K., Oenema, A., Prins, R., Brug, J., 2008. Socio-demographic factors as correlates of active commuting to school in Rotterdam, the Netherlands. Prev. Med. 47 (4), 412–416. https://doi.org/10.1016/j.ypmed.2008.06.019.
Berry, M., Kogan, J., 2010. Text Mining. John Wiley & Sons Ltd, United Kingdom.
Biehl, A., Chen, Y., Sanabria-Véaz, K., Uttal, D., Stathopoulos, A., 2019. Where does active travel fit within local community narratives of mobility space and place? Transp. Res. A 123, 269–287.
Blei, D.M., Lafferty, J.D., 2009. Topic Models. In: Srivastava, A., Sahami, M. (Eds.), Text Mining: Classification, Clustering and Applications. Chapman and Hall/CRC, Cambridge, pp. 71–93.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Buka, I., Koranteng, S., Osornio-Vargas, A.R., 2006. The effects of air pollution on the health of children. *Journal* of *Paediatrics and Child Health* 1111 (8), 513–516.
Cairns, S., Sloman, L., Newson, C., Anable, J., Kirkbride, A., Goodwin, P., 2004. Smarter Choices – Changing the Way We Travel. Department for Transport, London.

Cardoso, J.L., Stefan, C., Elvik, R., Sørensen, M. (2008) *Road Safety Inspection: Best Practice and Implementation Plan*. INCVC 3. LNEC, Lisboa ISBN 978-972-49-2138-9.
Carver, A., Timperio, A., Crawford, D., 2013. Parental chauffeurs: what drives their transport choice? J. Transp. Geogr. 26, 72–77. https://doi.org/10.1016/j.jtrangeo.2012.08.017.
Carver, A., Barr, A., Singh, A., Badland, H., Mavoa, S., Bentley, R., 2019. How are the built environment and household travel characteristics associated with children's active transport in Melbourne, Australia? J. Transp. Health 12, 115–129. https://doi.org/10.1016/j.jth.2019.01.003.
Chang, J., Boyd-Graber, J.L.., Gerrish, S., Wang, C., Blei, D.M. (2009) Reading tea leaves: how humans interpret topic models. Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada, pp. 288–296.
Cooper, A.R., Jago, R., Southward, E.F., Page, A.S., 2012. Active travel and physical activity across the school transition: the PEACH project. Med. Sci. Sports Exerc. 44 (10), 1890–1897.
Das, S., Sun, X., Dutta, A., 2010. Text mining and topic modeling of compendiums of papers from transportation research board annual meetings. Transportation Research Record: Journal of the Transportation Research Board 2552, 48–56.
Davison, K.K., Werder, J.L., Lawson, C.T., 2007. Children's Active Commuting to School: Current Knowledge and Future Directions. Prev. Chronic Dis. 5 (3), A 100.
Del Castillo, J.M., Benitez, F.G., 2012. A Methodology for Modeling and Identifying Users Satisfaction Issues in Public Transport Systems Based on Users Surveys. Procedia. Soc. Behav. Sci. 54, 1104–1114.
dell'Olio, L., Ibeas, A., Cecin, P., 2011. The quality of service desired by public transport users. Transp. Policy 18, 217–227.
DiMaggio, P., Nag, M., Blei, D., 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of US Government arts funding. Poetics 41 (6), 570–606.
Dyer, T., Lang, M., Stice-Lawrence, L., 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. J. Account. Econ. 64 (2–3), 221–245.
Eboli, L., Mazzula, G., 2009. A new customer satisfaction index for evaluating transit service quality. J. Public Transp. 12, 3.
Esztergár-Kiss, D., Tettamanti, T., 2019. Stakeholder engagement in mobility planning. Autonomous Vehicles and Future Mobility 113–123. https://doi.org/10.1016/b978-0-12-817696-2.00009-3.
Ewing, R., Schroeer, W., Greene, W., 2004. School location and student travel analysis of factors affecting mode choice. Transportation Research Record: Journal of the Transportation Research Board. 1895, 55–63. https://doi.org/10.3141/1895-08.
Faulkner, G.E.J., Richichi, V., Buliung, R.N., Fusco, C., Moola, F. (2010) What's "quickest and easiest?": parental decision making about school trip mode. *International Journal of Behavioral Nutrition Physical. Activity* 7:62 10.1186/1479-5868-7-62.
Feinerer, I., Hornik, K., Meyer, D., 2008. Text mining infrastructure in R. J. Stat. Softw. 25, 1–54.
Fowler, F.J.J. (1995) Improving survey questions. Design and Evaluation. Applied Social Research Methods Series.Vol.38. Sage Publications, USA.
Freeman, C., Quigg, R., 2009. Commuting lives: Children's mobility and energy use. J. Environ. Plan. Manag. 52 (3), 393–412. https://doi.org/10.1080/09640560802703280.
Fujii, S., Taniguchi, A., 2005. Reducing family car-use by providing travel advice or requesting behavioural plans: An experimental analysis of travel feedback programs. Transp. Res. Part D: Transp. Environ. 10 (5), 385–393. https://doi.org/10.1016/j.trd.2005.04.010.
Gao, L., Yu, Y., Liang, W., 2016. Public Transit Customer Satisfaction Dimensions Discovery from Online Reviews. Urban Rail Transit. 2, 146–152. https://doi.org/10.1007/s40864-016-0042-0.
Ghazizadeh, M., Mcdonald, A.D., Lee, J.D., Madison, W., 2014. Text Mining to DecipherFree-Response Consumer Complaints: Insights From the NHTSA Vehicle Owners' Complaint Database. Hum. Factors 56 (6), 1189–1203.
Griffiths, T., Steyvers, M., 2004. Finding scientific topics. PNAS PNAS), 101(Supplement 1, 5228–5235.
Grun, B., Hornik, K., 2011. Topicmodels: An R package for fitting topic models. J. Stat. Softw. 40 (30), 1–30.
Heelan, K.A., Abbey, B.M., Donnelly, J.E., Mayo, M.S., Welk, G.J., 2009. Evaluation of a walking school bus for promoting physical activity in youth. J. Phys. Act. Health 6 (5), 560–567.
Heinrich, G. (2005) Parameter estimation for text analysis. Technical report. URL http://www.arbylon. net/publications/text-est2.pdf. (accessed on: 2011-11-11).
Imob (2017) Inquérito à Mobilidade nas Áreas Metropolitanas do Porto e Lisboa. INE (Instituto Nacional de Estatística), Lisboa, Portugal.
INE (2021) Área Metropolitana de Lisboa in Figures. ISBN 978-989-25-0623-4, Lisboa, Portugal.
ITF Transport Outlook, 2017. The Organisation for Economic Co-operation and Development (OECD). Retrieved from: https://www.oecd.org/about/publishing/itf-transport-outlook-.
Iwata, T., Saito, K., Ueda, N., Stromsten, S., Griffiths, T., Tenenbaum, J., 2007. Parametric embedding for class visualization. Neural Comput. 19 (9), 2536–2556.
James, B., Burke, M., Yen, B.T., 2017. A critical appraisal of Individualised Marketing and Travel Blending interventions in Queensland and Western Australia from 1986–2011. Travel Behav. Soc. 8, 1–13. https://doi.org/10.1016/j.tbs.2017.03.002.
Johansson, K., Laflamme, L., Hasselberg, M., 2011. Active commuting to and from school among Swedish children–a national and regional study. Eur. J. Pub. Health 22 (2), 209–214. https://doi.org/10.1093/eurpub/ckr042.
Jones, A., Steinbach, R., Roberts, H., Goodman, A., Green, J., 2012. Rethinking passive transport: Bus fare exemptions and young people's wellbeing. Health Place 18, 605–612.
Karanasiou, A., Viana, M., Querol, X., Moreno, T., de Leeuw, F., 2014. Assessment of personal exposure to particulate air pollution during commuting in European cities-Recommendations and policy implications. Sci. Total Environ. 490, 785–797. https://doi.org/10.1016/j.scitotenv.2014.05.036.
Lee, N.R., Kotler, P., 2011. Social marketing: Influencing behaviors for good. SAGE publications, USA.
Long, K., Silva, D.C., Dias, F., Khoeini, S., Bhat, A.C., Pendyala, R.M., Bhat, C.R., 2019. Role of Childhood Context and Experience in Shaping Activity-Travel Choices in Adulthood. Transportation Research Record: Journal of the Transportation Research Board 2673. https://doi.org/10.1177/0361198119840338.
Mackay, R., 2005. The impact of family structure and family change on child outcomes: A personal reading of the research literature. Soc. Policy J. N. Z. 24, 111–133.
Manning, C.D., Schutze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, second edition, Cambridge, MA.
McCarthy, E.J., 1960. Basic marketing: A managerial approach. R.D. Irwin, USA, Homewood, III.
McDonald, N.C., 2008a. Household interactions and children's school travel: The effect of parental work patterns on walking and biking to school. J. Transp. Geogr. 16 (5), 324–331.
McDonald, N.C., 2008b. Children's mode choice for the school trip: The role of distance and school location in walking to school. Transportation 35 (1), 23–35.
McDonald, N.C., Steiner, R.L., Lee, C., Smith, T.R., Zhu, X., Yang, Y., 2014. Impact of the safe routes to school program on walking and bicycling. J. Am. Plann. Assoc. 80, 153–167. https://doi.org/10.1080/01944363.2014.956654.
McFarland, D.A., Ramage, D., Chuang, J., Heer, J., Manning, C.D., Jurafsky, D., 2013. Differentiating language usage through topic models. Poetics 41 (6), 607–625.
Mehrotra, S., and Roberts, S.C. (2018) Identification and validation of themes from vehicle owner complaints and fatality reports using text analysis. Proceedings of the Transportation Research Board 97th Annual Meeting 2018, Washington, D.C..
Mindell, J., Ergler, C., Hopkins, D., Mandic, S., 2021. Taking the bus? Barriers and facilitators for adolescent use of public buses to school. Travel Behav. Soc. 22, 48–58.
Mitra, R., Buliung, R.N., 2014. The influence of neighborhood environment and household travel interactions on school travel behaviour: an exploration using geographically-weighted models. J. Transp. Geogr. 36, 69–78. https://doi.org/10.1016/j.jtrangeo.2014.03.002.
Mouwen, A., 2015. Drivers of customer satisfaction with public transport services. Transp. Res. A Policy Pract. 78, 1–20.
Nikita, M. (2016) Tuning of the Latent Dirichlet Allocation Models Parameters. R Package Ldatuning Version 0.2.0. Comprehensive R Archive Network (CRAN).
National School Transportation Association, 2013. The Yellow School Bus Industry. Industry white paper, NSTA.
Panter, J.R., Jones, A.P., van Sluijs, E.M.F., 2008. Environmental determinants of active travel in youth: a review and framework for future research. International Journal Behavioral Nutrition Physical Activity 5, 1–14. https://doi.org/10.1186/1479-5868-5-34.
Papaionnou, D., 2017. Assessing the realtion between mode choice, user satisfaction, and quality for Public Transport systems. Instituto Superior Técnico, Universidade de Lisboa, Portugal. Ph.D. Thesis,.

Pont, K., Ziviani, J., Wadley, D., Abbott, R., 2011. The model of children's active travel (M-CAT): a conceptual framework for examining factors influencing children's active travel. Aust. Occup. Ther. J. 58, 138–144. https://doi.org/10.1111/j.1440-1630. 2010.00865.x.

Preston, J., 2012. Integration for Seamless Transport. ITF Discussion Paper 2012–01. OECD/ ITF 01 (1), 1–34. https://doi.org/10.1787/5k8zvv8lmswl-en.

Queiroz, M.M., Celeste, P., Moura, F. (2019) School commuting: the influence of soft and hard factors to shift to public transport. 22nd EURO Working Group on Transportation Meeting, EWGT2019 Spain.

Queiroz, M.M., Celeste, P., Moura, F. (2020a) Matching users' expectations in school public behavior: where are we in public transport?. Proceedings of Transport Research Arena, TRA 2020, Helsinki, Finland.

Queiroz, M.M., Roque, C., Moura, F., 2020b. Shifting from Private to Public Transport using a Duration-Based Modeling of a School-Based Intervention. Transportation Research Record: Journal of the Transportation Research Board 2674 (7), 540–554. https://doi.org/10.1177/0361198120923666.

Ramos, S., Vicente, P., Passos, A.M., Costa, P., Reis, E., 2019. Perceptions of the Public Transport Service as a Barrier to the Adoption of Public Transport: A Qualitative Study. Soc. Sci. 8 (5), 150. https://doi.org/10.3390/socsci8050150.

Roberts, S.C. and Lee, J.D. (2014) Deciphering 140 Characters: Text Mining Tweets On #DriverDistraction. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2195–2199.

Robertson-Wilson, J.E., Leatherdale, S.T., Wong, S.L., 2008. Social-Ecological Correlates of Active Commuting to School Among High School Students. J. Adolesc. Health 42 (5), 486–495. https://doi.org/10.1016/j.jadohealth.2007.10.006.

Roque, C., Cardoso, J.L., Connell, T., Weber, R. (2019) Topic analysis of Road safety inspections using Latent Dirichlet Allocation: A case study of roadside safety in Irish main roads. Accident Analysis and Prevention. doi.org/10.1016/j.aap.2019.07.021.

Sahlqvist, S., Veitch, J., Abbott, G., Salmon, J., Garrard, J., Acker, F., Hartman, K., Timperio, A., 2019. Impact of an Australian state-wide active travel campaign targeting primary schools. Prev. Med. Rep. 14, 100866 https://doi.org/10.1016/j.pmedr.2019.100866.

Sauvage-Mar, C., Naylor, P.-J., Higgins, J., VonBuchholz, H., 2019. Way2Go! Social marketing for girls' active transportation to school. Prev. Med. Rep. 100828 https://doi.org/10.1016/j.pmedr.2019.100828.

Selby, E.A., Wonderlich, S.A., Crosby, R.D., Engel, S.G., Panza, E., Mitchell, J.E., Crow, S., Peterson, C.B., Le Grange, D., 2014. Nothing tastes as good as think feels: low positive emotion differentiation and weight-loss activities in anorexia nervosa. *Clinical Psychological. Sci*Ence 2, 514–531. https://doi.org/10.1177/2167702613512794.

Stark, J., Singleton, P.A., Uhlmann, T., 2019. Exploring children's school travel, psychological well-being, and travel-related attitudes: Evidence from primary and secondary school children in Vienna, Austria. Travel Behav. Soc. 16, 118–130. https://doi.org/10.1016/j.tbs.2019.05.001.

Sun, L., Yin, Y., 2017. Discovering themes and trends in transportation research using topic modeling. Transportation Research Part c: Emerging Technologies 77, 49–66. https://doi.org/10.1016/j.trc.2017.01.013.

Timperio, A., Ball, K., Salmon, J., Roberts, R., Giles-Corti, B., Simmons, D., Baur, L.A., Crawford, D. (2006) Personal, family, social, and environmental correlates of active commuting to school. *American Journal of Preventive Medicine* 30(1):45–51. PubMed doi:10.1016/j.amepre.2005.08.047.

Tudor-Locke, C., Ainsworth, B.E., Popkin, B.M., 2001. Active commuting to school: An overlooked source of children's physical activity? Sports Med. 31 (5), 309–313.

United Nations, 2016. Climate change. Retrieved from: The Paris Agreement. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.

Valença, G., Moura, F., Morais de Sá, A., 2023. Where is it complex to reallocate road space? Environment and Planning B: Urban Analytics and City. Science 3(1). https://doi.org/10.1016/j.jjimei.2022.100150.

Wang, Y., Sabzmeydani, P., Mori, G. (2007) Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. Proceedings of Human Motion - Understanding, Modeling, Capture and Animation, Second Workshop, Human Motion, Rio de Janeiro, Brazil.

Waygood, E.O.D., Friman, M., Olsson, L.E., Taniguchi, A., 2017. Transport and child well-being: An integrative review. Travel Behav. Soc. 9, 32–49.

Westman, J., Olsson, L.E., Garling, T., Friman, M., 2017. Children's travel to school: satisfaction, current mood, and cognitive performance. Transportation 44 (6), 1365–1382. https://doi.org/10.1007/s11116-016-9705-7.

Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., Zou, W., 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinf*Ormatics 16 (13), S8.

Zhu, X.M., Lee, C., 2009. Correlates of Walking to School and Implications for Public Policies: Survey Results from Parents of Elementary School Children in Austin, Texas. J. Public Health Policy. 30:S177–S202. PubMed. doi:10.1057/jphp.2008.51.