Accident Analysis and Prevention 131 (2019) 336–349

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Topic analysis of Road safety inspections using latent dirichlet allocation: A case study of roadside safety in Irish main roads

Carlos Roque[a,*], João Lourenço Cardoso[a], Thomas Connell[b], Govert Schermers[c], Roland Weber[d]

[a] Laboratório Nacional de Engenharia Civil, Departamento de Transportes, Núcleo de Planeamento, Tráfego e Segurança, Av do Brasil 101, 1700-066 Lisboa, Portugal
[b] Arup, 50 Ringsend Road, Dublin, D04 T6X0, Ireland
[c] SWOV Institute for Road Safety Research, Bezuidenhoutseweg 62, 2509 AC The Hague, the Netherlands
[d] RW, Germany

ARTICLE INFO

Keywords:
Roadside safety
Road Safety Inspection
Text mining
Topic Modeling
Latent Dirichlet allocation

ABSTRACT

Under the Safe System framework, Road Authorities have a responsibility to deliver inherently safe roads and streets. Addressing this problem depends on knowledge of the road network safety conditions and the number of funds available for new road safety interventions. It also requires the prioritisation of the various interventions that may generate benefits, increasing safety, while ensuring that reasonable steps are taken to remedy the deficiencies detected within a reasonable timeframe. In this context, Road Safety Inspections (RSI) are a proactive tool for identifying safety issues, consisting of a regular, systematic, on-site inspection of existing roads, covering the whole road network, carried out by trained safety expert teams.

This paper aims to describe how topic modelling can be effectively used to identify co-occurrence patterns of attributes related to the run-off-road crashes, as well as the corresponding patterns of road safety interventions, as described in the RSI reports. We apply latent Dirichlet allocation (LDA), a widespread method for fitting a topic model, to analyse the topics mentioned in RSI reports, divided into two groups: problems found; and proposed solutions. For this study, 54 RSI gathered over six years (2012–2017) were analysed, covering 4011 km of Irish roads.

The results indicate that important keywords relating to the "forgiving roadside" and "clear zone" concepts, as well as the relevant European technical standards (CEN-EN1317 and EN 12,767), are absent from the extracted latent topics. We also found that the frequency of topics related to roadside safety is higher in the problems record set than in the solutions record set, meaning that problems are more easily identified and related to the roadside area than interventions may be.

This paper presents methodological empirical evidence that the LDA is appropriate for identifying the co-occurrence patterns of attributes related to the ROR crashes in road safety inspections' reports, as well as the interventions' patterns associated with these crashes. Also, it provides valuable information aimed to determine the extent to which national road authorities in Europe and their contractors are currently capable of implementing and maintaining compliance with roadside standards and guidelines throughout the life cycle of roads.

1. Introduction

Under the Safe System framework, Road Authorities have a responsibility to deliver inherently safe roads and streets (ITF, 2016). Addressing this problem depends on knowledge of the road network safety conditions and the amount of funds available for new road safety interventions. It also requires the prioritisation of the various interventions that may generate benefits, increasing safety, while ensuring that judicious steps are taken to remedy the deficiencies detected within a reasonable timeframe. This process of intervention in road safety is continuous and recursive, and encompasses it the entire life cycle of roads – concept, planning, design, build, operate, maintain and decommission.

Infrastructure related road safety problems may be detected

* Corresponding author at: Laboratório Nacional de Engenharia Civil, Departamento de Transportes, Núcleo de Planeamento, Tráfego e Segurança, Av do Brasil 101, 1700-066 Lisboa, Portugal. Tel.: (+351) 218-443-970.
E-mail addresses: croque@lnec.pt (C. Roque), jpcardoso@lnec.pt (J. Lourenço Cardoso), Thomas.Connell@arup.com (T. Connell), govert.schermers@swov.nl (G. Schermers), rolandweber@gmx.de (R. Weber).

https://doi.org/10.1016/j.aap.2019.07.021
Received 24 April 2019; Received in revised form 17 July 2019; Accepted 20 July 2019
Available online 01 August 2019
0001-4575/

proactively (i.e., before crashes occur, also known as *a priori*) or reactively (i.e., after numerous crashes occur, or *a posteriori*), depending on whether site-specific data on crashes are explicitly required or not. A balanced road infrastructure safety management (RISM) system contains both forms of diagnosis and includes interventions at a network level, on selected routes or at specific locations. According to Hauer (1998), when applying a rational approach to such a system, the road safety manager aims at the efficient mitigation of the road crash burden through a selection of interventions that seeks to anticipate the consequences of decisions and interventions, to balance costs with benefits and which takes advantage of lessons learned from experience.

With this general setting, European Union (EU) Member States agreed under Directive 2008/96/EC (EU, 2008) to implement a Road Infrastructure Safety Management (RISM) system consisting of five supporting procedures:

- Road Safety Impact Assessment (RIA)
- Road Safety Audit (RSA)
- Black spot treatment (BST)
- Network Safety Management (NSM)
- Road Safety Inspection (RSI)

At the planning stage, a RIA is performed to assess the impact on the safety of a future road investment. This can be a new bridge that may or may not be intended to raise the network safety level; or the assessment of a wider scheme i.e. the plans for upgrading the safety level of a total network or area (Eenink et al., 2007). At the design stage, an RSA is carried out to ensure that new road schemes will operate as safely as possible for all road user groups. RSA consists of the examination of road schemes at the different stages of project development (starting with the preliminary design) until before or shortly after a road is opened to traffic (Proctor et al., 2003; Matena et al., 2007). Both RIA and RSA are proactive approaches.

Once fully operational, the safety level of an existing road may be improved through several types of procedures, including treatment of hazardous locations, network safety management, and road safety inspections. Black spot treatment consists of the identification, analysis, and treatment of black spots, which are defined as any location that has a higher number of crashes than other similar sites as a result of local risk factors. Network safety management is the identification, analysis, and treatment of hazardous road sections, which are defined as any section that has a higher expected number and severity of crashes than other comparable road sections, as a result of local and section based accident and injury factors (Sørensen and Elvik, 2007). Network safety management differs from black spot treatment by focusing on longer road sections of typically two to 10 km, while the black spots seldom are longer than 0.5 km. Both are reactive procedures.

RSI are a proactive tool for detecting safety issues on existing operating roads, consisting of a regular, systematic, on-site inspection, covering the whole road network, carried out by trained safety expert teams. Road hazards and safety issues detected with this activity are described in a written report, for which a formal response by the relevant road authority is required (Cardoso et al., 2008).

Developments in the road network may create a conflict between the current function of a road and its original intended use, along with the inadequacy of equipment and design characteristics to the current use of the road. Furthermore, improvements in road standards may result in discrepancies between characteristics of newly built or reconstructed roads and existing ones, interfering with the establishment of common *a priori* expectations concerning road use. Due to technological developments and new technical standards, existing road equipment may become obsolete, its replacement being necessary. These and others are hazardous factors emerging during the lifecycle of a road itinerary and generally unforeseeable during the planning and design stages. Tackling these hazards to raise the safety level of existing roads and bring their standards to adequate consistency with the rest of the road network is the main objective of RSI. A complementary outcome may also be achieved by RSI: to maintain or restore the original safety level of an existing road. However, it is recognized that most issues related to this secondary outcome should be mainly achieved through regular and more frequent road maintenance inspections.

RSI is a proactive tool for several road operators since its application to an itinerary or road section is not dependent on knowledge concerning the road specific safety level. Neither the decision for the initiation of a RSI nor the procedures for its execution require knowledge on the registered safety record of the relevant itinerary. To carry out a RSI, only general knowledge on road hazards, on safety issues related to the road environment and on effective infrastructure interventions are needed. However, some road agencies decided to also assess prospective benefits of proposed interventions using historic crash data (Cardoso et al., 2008).

The elements to be addressed in these RSI are known crashes or injuries risk factors, such as the quality of traffic signs, road markings and road surface characteristics; the adequacy of sight distances; the presence of roadside traffic hazards; and consistency between road function and key aspects of traffic operation (Cardoso et al., 2007).

Many EU member countries have already adopted the RSI procedure. An example is Ireland, where RSI was introduced by the Transport Infrastructure Ireland (TII) to comply with Directive 2008/96/EC, under TII AM-STY-06044 Road Safety Inspection, part of the Irish Design Manual for Roads and Bridges (Transport Infrastructure Ireland, 2014). In Ireland, RSI have been carried out on all national roads, since 2012.

An individual RSI report is prepared by each Irish RSI team for each inspected route, containing a brief description of each safety issue (hazard), an informal risk assessment of each identified hazard, and recommendations to deal with it. The final report is submitted to TII (Transport Infrastructure Ireland, 2014). These RSI reports have a standardized structure, and the descriptive analysis they follow has a pre-set methodology. The large number of issues described in the RSI, and the detailed description of each item, yields a large-scale dataset, which may support identifying patterns of common conditions and related interventions. Roadside safety can be improved by identifying Run-off-road (ROR) crash conditions that tend to co-occur in these events, followed by evidence-based roadside safety interventions. However, the reports are generally narrative (written), and thus, it is difficult to extract hazard patterns from a large set of reports; the same happens with corrective intervention patterns. With this framework, the Irish situation was selected as a case study to determine whether a methodology could be applied with which to extract re-occurring or common problems and corrective measures. Thus, the focus of this study is not in analysing the locations where problems were detected, and interventions proposed, but to identify recurring patterns in RSI results and to assess their correspondence with safe system principles.

This paper aims to describe how topic modelling can be effectively used to identify co-occurrence patterns of attributes related to the ROR crashes, as well as the corresponding patterns of road safety interventions, as described in RSI reports, using those produced for Transport Infrastructure Ireland (TII) in Ireland as an example. This study was done within the CEDR research project PROGReSS – Provision of Guidelines for Road Side Safety which was funded under the 2016 Safety Call. In this project, the results of a status quo review of available EU roadside safety standards and guidelines are combined with the experiences from National Road Authorities in applying these in the design, operation and maintenance phases of EU high-speed roads (with speed limits higher than 70 km/h). A particular emphasis is put on the six funding countries (Belgium-Flanders, Ireland, Netherlands, Slovenia, Sweden, United Kingdom), plus Germany and Portugal which are included to increase the geographic representation of the results. For this study, 54 RSI reports produced in Ireland and submitted to TII over six years (2012 to 2017) were analysed. Other sources of information produced by road authorities during road inspections that

might have been used in this study include e-mail communications, reports to web pages and social media content. The analysed RSI reports comprise a significant portion of unstructured content in textual format.

One possibility to analyse this huge amount of information is text mining, which, according to Canito et al. (2018), is a tool for extracting information from textual data, which can then be used for research or business purposes. Within PROGReSS a specific data driven approach was developed in order to identify many-to-many[1] associations among a broad group of conditions associated with ROR crashes and roadside safety interventions.

In this paper, the following terms are used, as defined by Blei et al. (2003):

•A *Word* or *Term* is the basic unit of discrete data, defined to be an item from a vocabulary indexed by {1,…, V};

•A *Document* is a sequence of N *words* denoted by $F = (w_1, w_2,\ldots, w_N)$, where w_n is the n^{th} word in the sequence;

•A *Corpus* is a collection of M *documents* denoted by $D = (F_1, F_2,\ldots, F_N)$;

•A *Topic* is a unique distribution over a vocabulary of words that inspection teams use to identify road safety issues and to describe their proposed solutions.

Considering the technical requirements of text analysis, this research applies latent Dirichlet allocation (LDA) (Blei et al., 2003), a particularly common method for fitting a topic model, to analyse the topics of RSI, divided into two groups: problems found; and proposed solutions. A topic model allows one to algorithmically identify *topics* within a collection of *documents* based on the *words* contained in each *document* (Ghosh and Guha, 2013). The LDA algorithm is a three-level hierarchical Bayesian modelling process, which groups a set of items into topics defined by *words* or *terms*, where each of the *terms* identified characterizes a *topic* (Blei et al., 2003).

Underlying the "*bag-of-words*" assumption (i.e., the order of words in a document can be neglected) LDA represents a *document* as a mixture of latent topics in which a topic has a multinomial distribution over words. Every *document* will have its own mixing proportion of topics, and each topic has its own *word* distribution (Wang et al., 2018).

Based on an unsupervised Bayesian learning algorithm, LDA can capture the latent topics that represent the opinions of the inspection teams from unstructured and large written reports. Each topic can be regarded as a specific feature of the issue or road reported by inspection team members that represents aspects which are frequently mentioned in unstructured large written reports developed by these teams.

In this case, LDA was applied to two datasets collected from the RSI reports, one comprising the issues raised (detected dangers), and the other the proposed solutions. In both cases, informative, distinct and tight topics were obtained, which aligned well with known co-occurrences among conditions cited in the literature.

The method's ability to generate meaningful topics from both datasets, where one comprises more common dangerous conditions (issues raised) and the other safety interventions (proposed solutions), demonstrates its effectiveness in reliably exposing co-occurring attributes.

Notably, the results uncover a few indirect associations among conditions that have previously gone unreported, suggesting that topic modelling over RSI reports can expose yet unnoticed associations, in this case contributing to improved knowledge on roadside safety issues.

2. Methodology

Linguistics and computer science together advanced the empirical study of language in the field of computational linguistics and natural language. This field has developed a suite of methods capable of identifying patterns of language usage in large bodies of text and communication. These methods include supervised and unsupervised document classification techniques. The first can be used to automatically assign one of an existing set of labels to new documents, where some external mechanism, such as personal feedback, provides information on the correct classification for documents. The latter can place related documents together, based on the words they possess without using external label information processing (McFarland et al., 2013). Both unsupervised and supervised topic models have been recently applied to examine language in road safety (Pereira et al., 2013; Bao et al., 2018; Qi and Guan, 2019; Zhang et al., 2018).

2.1. Text mining

Word combinations are an important source of information (Arnon and Snider, 2010). The relationships between two words can be analysed by counting how often word X is followed by word Y. Two-word phrases (hereafter bigrams) are more informative than individual words as they provide some degree of context (Ghosh and Guha, 2013). By automatically extracting and using phrases, especially bigrams, it is possible to improve the identification of the road safety issues and interventions described in the RSI reports.

The relationships between two words can also be discovered through correlation analysis. Correlation analysis seeks to expose the relationships between words that are found in the same *document*, but may not co-occur such as with bigrams (Silge and Robinson, 2017). In this case, the focus is to determine how often words appear together in the same document (previously defined as a sequence of *N* words) relative to how often they appear separately. To implement this formulation, the mean square contingency coefficient (also known as phi coefficient) was calculated. The mean square contingency coefficient measures the extent and direction of correlation between two variables (Selby et al., 2004).

The most basic text mining approach is the tf-idf scheme (Salton and McGill, 1983), which counts the number of occurrences of each word for each document and applies appropriate normalization on the frequencies of the words. The result of tf-idf is a document-term matrix whose columns contain the normalized frequency of the words for each of the documents in the analysed corpus. According to Bastani et al. (2019), this approach it is not helpful where the number of unique words is very large, even if it reduces documents of arbitrary length into a fixed-length of the unique words in the corpus. Deerwester et al. (1990) proposed latent semantic analysis (LSA), which is an effective dimensionality reduction method on the document-term matrix using singular value decomposition. This approach can achieve significant compression in large collections while capturing some aspects of basic linguistic notions. However, the main shortcoming of LSA is the lack of fitting a model to data to represent the documents into multiple topics (Bastani et al., 2019).

A significant improvement in this regard was initiated with probabilistic Latent Semantic Indexing (pLSI) from Hofmann (1999). The pLSI approach assigns a probabilistic mixture model to the words in a document, where the mixture components represent topics. However, pLSI is prone to overfitting. Overfitting is a phenomenon where almost perfect prediction can be made by a classifier on a training data set, while achieving poor prediction on test data. This can lead to fitting the noise in the data by learning the details of the training data instead of finding a global predictive formula (Hamed et al., 2018). LDA has been introduced to solve overfitting in the original pLSI (probabilistic Latent Semantic Indexing) (Ihou and Bouguila, 2019).

Recently, several modifications to LDA have been proposed to incorporate supervision and enable categorization schemes. One form of supervision assumes that a single label of metadata is generated from each document's empirical topic mixture distribution, such as in Supervised LDA (McAuliffe and Blei, 2008). Roberts et al. (2014)

[1] In many-to-many associations each object can be related to multiple objects of another type. For example, a doctor can have many patients, and a patient can also have many doctors.

developed an extension of LDA, known as structural topic modeling (STM), to take the effect of covariates of interest in documents modelling. Also, advances in nonparametric Bayesian methods resulted in topic models that use nonparametric Bayesian priors, like the hierarchical Dirichlet process LDA (HDP-LDA) proposed by Teh et al. (2006) that can estimate the topic and word priors and to infer the number of clusters from the data.

2.2. Topic modelling

A method suited to the study of high-level relationships between documents is a class of probabilistic techniques called "topic models", such as the LDA (Blei et al., 2003) applied in this study, which identifies distinct "bags" of words co-occurring in documents. LDA is called a "topic model" because the identified sets of words tend to reflect underlying topics that may be combined to characterize every document in a corpus (Blei et al., 2003). Topic modelling is a method that models each document as a mixture of topics and each topic as a mixture of words.

LDA is an unsupervised topic model where the number of topics that the model discovers is left as a free parameter. Being unsupervised, LDA does not incorporate manual notation into the learning procedure of topics. Common to all unsupervised topic models is the idea that language is organized by latent dimensions that actors may not even be aware of (McFarland et al., 2013).

In this study, LDA was employed to model problem and solution documents as though they were generated by sampling from a mixture of K topics, where a topic is a multinomial distribution over all words in our vocabulary (Blei et al., 2003).

The generative process for each file of problems or solutions consists of the following steps (Bhattacharya et al., 2018):

- First, a multinomial distribution over V words for the t^{th} topic, denoted Φ_t $(1 \leq t \leq K)$, is obtained by sampling from a Dirichlet distribution with parameter α; Φ_t representing the conditional probability of a word to occur in the t^{th} topic.
- Next, for each document, F_i, a multinomial distribution over K topics, denoted θ_i, is sampled from a Dirichlet distribution with parameter β; θ_i represents the conditional probability of the file to be associated with each of the K topics.
- Subsequently, for each word-position, j, in the document, F_i:
  - (1) A topic is drawn by sampling from θ_i; the selected topic at position j in F_i, is denoted $z_j^i \in \{1,\ldots,K\}$;
  - (2) Given the topic z_j^i a word c_j^i is drawn by sampling the topic-word distribution, $\theta_{z_j^i}$.

The model parameters are set iteratively for different values of K (in our study K ranges from two to 25), and the data log-likelihood is calculated for each value of K. To determine the optimal number of topics, we identify the K value that maximizes the data log-likelihood, which is defined as:

$$\sum_{i=1}^{M} log \int \left\{ \sum_{z} \left[\sum_{j=1}^{N_j} Pr(c_j^i | z_j^i, \theta_{z_j^i}) Pr(z_j^i | \theta_i) \right] \right\} Pr(\theta_i | \beta) d\theta_i \quad (1)$$

where M denotes the number of documents in the corpus and N_i denotes the total number of words in the i^{th} file. Further details may be obtained in Grun and Hornik (2011).

The exact parameter inference of the LDA model is intractable, and thus, approximate estimation methods are needed. The approximate algorithm Gibbs sampling (Griffiths and Steyvers, 2004) is widely used for parameter estimation in topic models due to its simplicity under Dirichlet priors (Wang et al., 2018).

Since the extracted topics, expressed as collections of words, are inherently latent, they often contain multi-dimensional meanings (semantics).

A parallel understanding can be drawn between topics in LDA and principal components in Principal Component Analysis (PCA). LDA turns a text document (represented by word frequencies) into a linear combination of topics (also represented by word frequencies). This linear combination of topics are similar to the eigenvectors in PCA. In PCA, a numerical signal of dimension N can be re-represented by a combination of K eigenvectors $K < N$. These dimensionality reductions lead to a loss of information (unless $K = N$), and the modeller has to seek the best compromise. In the case of LDA, given a set of documents (represented as a vector of word frequencies) and a number of K topics, the algorithm extracts the set of K topics that minimizes the reconstruction error of the original documents. Each topic is also a vector of word frequencies (Li et al., 2015).

This characteristic of LDA is useful to efficiently extract from RSI reports those words and documents which are specifically related to roadside safety.

To perform the text mining procedure, the statistical open-source tool R Version 3.4.2 (R Development Core Team, 2011) was adopted. Specifically, the "tm" (Feinerer et al., 2008) and "topicmodel" packages (Grun and Hornik, 2011) were chosen. The former provides text mining functions, while the latter implements the LDA algorithm.

The R "topicmodel" package (Grun and Hornik, 2011) uses Gibbs sampling. Estimation of the LDA model using Gibbs sampling requires specification of values for the parameters of the prior distributions. Griffiths and Steyvers (2004) suggest an initial value of 50/M for α and 0.1 for β. The parameter values used for the parameter β (0.1) and for the initial value of the parameter α (50/M) were suggested by Griffiths and Steyvers (2004).

The number of latent topics (K) to be estimated by the LDA algorithm is a problem of model selection, where the number of topics within a given corpus is initially unknown and must be chosen before model initialization. While LDA uses Bayesian inference to generatively estimate the posterior model distribution based only on the words shown in the texts, it requires K to begin its iteration process.

Since LDA is an unsupervised method, there is no direct measure to identify the optimum number of topics to include in a model (Robinson, 2019). Researchers have recommended various approaches to establish the optimal K (Arun et al., 2010; Cao et al., 2009; Deveaud et al., 2014; Griffiths and Steyvers, 2004; Zhao et al., 2015). These approaches provide a good range of possible K values that are mathematically plausible. However, according to DiMaggio et al. (2013), when topic modelling is used to identify themes and assist in interpretation (like in the present study), rather than to predict a knowable state or quantity, there is no statistical test for the optimal number of topics or the quality of a solution. A simple way to evaluate topic models is to look at the qualities of each topic and discern whether they are reasonable (McFarland et al., 2013).

In this paper, two different approaches were used to establish the optimum number of topics:

- KL-divergence minimization method (Arun et al., 2010);
- Expectation maximization method (Griffiths and Steyvers, 2004).

The R package "ldatuning" (Nikita, 2016) was used to set the optimal K.

In addition, the topic number selection was guided by the model's ability to identify a number of substantively meaningful and analytically useful topics. In fact, the increase in fit is sometimes at the expense of interpretability due to overfitting (Dyer et al., 2017). Increasing the number of topics, producing ever-finer partitions can result in a less useful model because it becomes almost impossible for humans to differentiate between many of the topics (Chang et al., 2009). Ultimately, the choice of models must be driven by the questions being analysed. DiMaggio et al. (2013) suggest that the process is empirically disciplined, in that, if the data are inappropriate for answering the analysts' questions, no topic model will produce a useful reduction of the

Table 1
Summary description of inspected Irish routes length (km).

Urban	Rural	Motorway	Single Carriageway	Dual Carriageway	National Primary	National Secondary	Total
626	3361	24	3846	141	1494	2507	4011

data.

3. Data

The main element of a RSI is the identification of the road safety issues and associated risks, in this case from a set of RSI reports generated by the National Road Authority in Ireland (Transport Infrastructure Ireland, 2014). Each RSI report includes attributes such as identification number, whether the issues occur on the mainline or on the roadside, a detailed description of the safety issues, the related primary collision type, the recommended approach to address each safety issue, and detailed descriptions of proposed solutions to eliminate the safety issues or mitigate their consequences. Each RSI report also includes an appendix (Appendix A) with a summary spreadsheet of all issues. A sample report format is contained in the Irish guidance document NRA HA 17 Road Safety Inspection Guidelines (Transport Infrastructure Ireland, 2014).

Table 1 shows a summary description of the routes analysed in the 54 investigated RSI.

The topic analyses focused mainly on the spreadsheets containing the summary of identified safety problems and the corresponding proposed mitigating interventions (Appendix A of the reports). Furthermore, the content of the inspection teams' reports, including the description of safety issues and proposed solutions were tracked, and two record sets – document-term matrices were constructed.

To create a document-term matrix that can be processed via topic modelling several data organization and pre-processing choices were made. The document-term matrix serves as input to the LDA topic modelling to obtain the most relevant topics (Blei et al., 2003).

Text pre-processing in this study includes word text tokenization (breaking up a sequence of strings into tokens such as words), converting words to lower-case, removing punctuation characters and numbers, and removing stop words (highly frequent words contributing little or no meaning in the text, such as "if", "and", etc.).

Stemming (reducing inflected words to their base or root form) was not considered in pre-processing, since it sometimes combines terms that would best be considered distinct, and variations of the same word will usually end up in the same topic.

Fig. 1 shows descriptions of the most frequent words appearing in each record set, in decreasing order of frequency occurrence. Each bar represents the number of occurrences of each word in the respective record set. Examining the words within each record set shows that words related to roadside issues are part of the most frequent words present in both record sets. Words like "control", "loss", "collision", "errant", "hazard", in the problems record set, or "side", "barrier", "post", "relocate" and "remove", in the solutions record set, point out to roadside safety issues that can be further examined through topic modelling.

4. Results and discussion

4.1. Relationships between words

Figs. 2 and 3 present a combination of connected nodes for both the problems and the solutions record sets, respectively, where it is possible to visualize some details of the text structure. The relationships here are directional (marked with an arrow).

In Fig. 2 one can see that words such as "road", "vehicle", "sign", and "roadside" form common centres of nodes. The word "roadside" is preceded by "unforgiving" and followed by "hazard". We also see pairs and triplets that form common short phrases related to roadside issues ("safety barrier", "bridge parapet" or "errant vehicle enters/striking"). Fig. 2 also shows the more general character of RSI, highlighting visibility and sight distance problems associated with road, vehicle and the bigram "drivers-inappropriately".

Fig. 3 shows that the solutions record set is particularly focused around words such as "sign", "signs" and "signage". In terms of roadside interventions, the phrases "safety barrier board" and "roadside boundary wall" stand out. Similar to the previous figure, Fig. 3 also highlights the broader set of remedial actions in RSI (e.g., vulnerable road users and layout review).

In Figs. 4 and 5, the correlations (through the mean square contingency coefficients) are depicted among words for the problems and solutions record sets, respectively. Note that unlike the bigram analysis, the relationships here are symmetrical, rather than directional. It can also be seen that while pairings of words that dominate bigram pairings are common, such as "unforgiving/roadside" or "barrier/safety", pairings of words that appear close to each other are also present, such as "clear" and "zone", "working" and "barrier", "utility" and "pole", or "forgiving" and "fence". The word "kerb" also appears correlated with several words (e.g. "provision", "dropped" and "paving").

It is worth mentioning that aquaplaning stands out as an issue in non-motorway RSI. Some correlations show the broader scope of RSI (e.g. "sight", "distance" and mainline", in Fig. 4; and "drainage", "setting" and "need", in Fig. 5). Vulnerable road users ("vru") and "surfacing" are at the core of two important correlated nodes in the solutions record set depicted in Fig. 5.

Compared with the most frequent words appearing in each record set (Fig. 1), the results extracted by directed graphs of common bigrams (Figs. 2 and 3) and correlations (Figs. 4 and 5) have better representation for the given RSI reports, including for roadside issues. Important words for roadside safety like "vru", "kerb", or "unforgiving" are displayed in Figs. 2 to 5 although they are not part of the most frequent words presented in Fig. 1.

Finally, in Fig. 6, the words most correlated with "barrier", "pole", "roadside", and "zone" are presented for the problems and solutions record sets. The word "barrier" is highly correlated with the words "working" and "width", and "terminal" and "P4″ in the problems and the solutions record sets, respectively. These correlations point out issues regarding the working width of installed safety barriers, as well as the need for the installation and selection of terminals at the extremities of highway safety barriers – "P4″ are guardrail end terminals designed for use on roads with speed limits of 50 mph (80 km/h) and greater. Also, examining the words correlated with the word "pole", it is possible to identify issues regarding utility poles and the need for their relocation" (through the words "utility" and "relocate", respectively). Similarly, the words "roadside" and "zone" are highly correlated with the words "unforgiving" and "clear", showing issues regarding the adoption and application of the "forgiving roadside" and "clear zone" concepts.

4.2. Models and interpretation

The LDA implementation was applied to both problems and solutions *corpora*, where each of the resulting topics is a distribution over words. Different numbers of topics, K, were considered, ranging from two to 25. To avoid the use of poor initial estimates as part of the Gibbs sampling process, 4000 samples were discarded in the burn-in period –

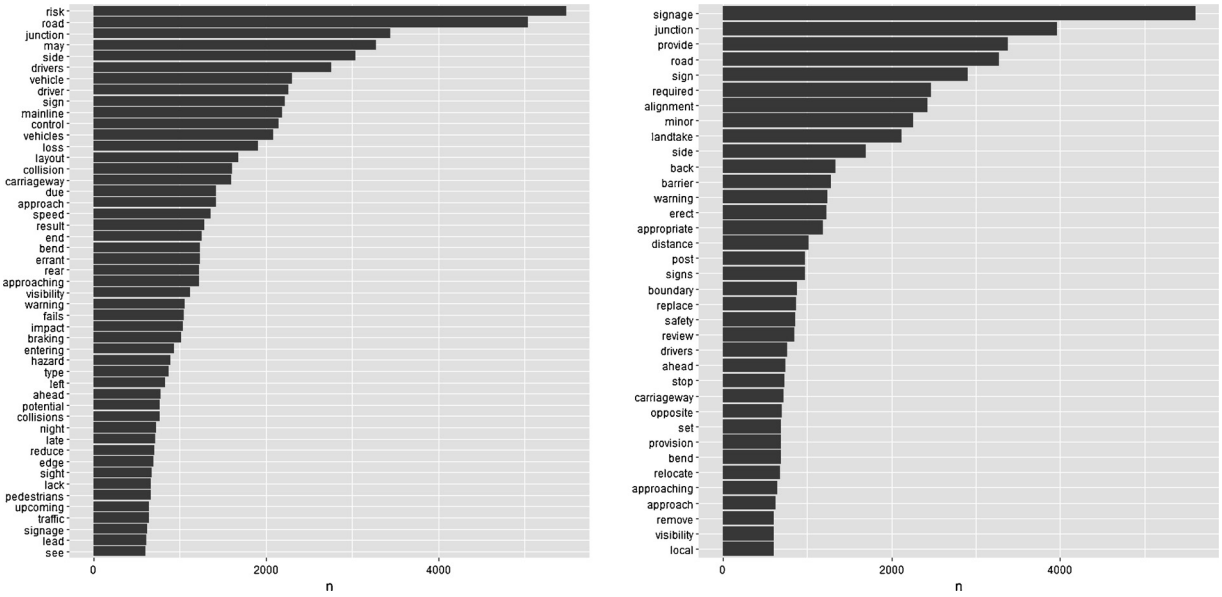

Fig. 1. Number of occurrences of the most frequent words in the problems record set (left), and the solutions record set (right).

the initial stage of the sampling process in which the Gibbs samples are poor estimates of the posterior (Bhattacharya et al., 2018). Following the burn-in period, 2000 iterations were performed, taking every 500th iteration for further use. This procedure is done to avoid correlations between samples. Each experiment was repeated five times employing different initial seeds and calculated an average log-likelihood value. The initial seeds were saved so that the results can be reproduced.

As shown in Fig. 7, the KL-divergence minimization method (top charts) and the expectation maximization method (bottom charts) agree that the ideal number of topics for our sample dataset is 25. Consequently, two LDA models were estimated by setting the K value equal to 25.

Tables 2 and 3 show the 25 extracted latent topics for the problems and solutions record sets (topics directly related to roadside issues are shaded grey). Each topic contains all words in the *corpus*, albeit with different probabilities. The top 10 terms for each record set are listed in

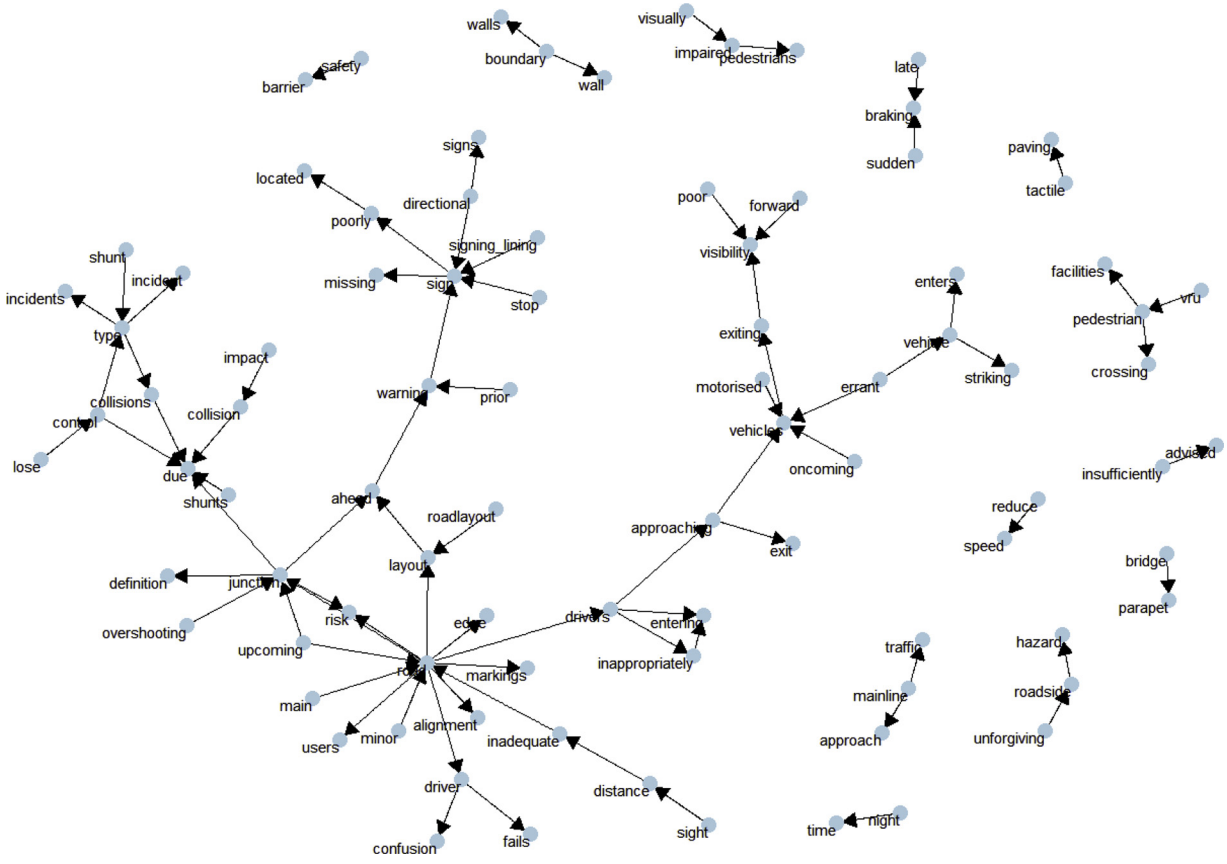

Fig. 2. Directed graph of common bigrams in the problems record set.

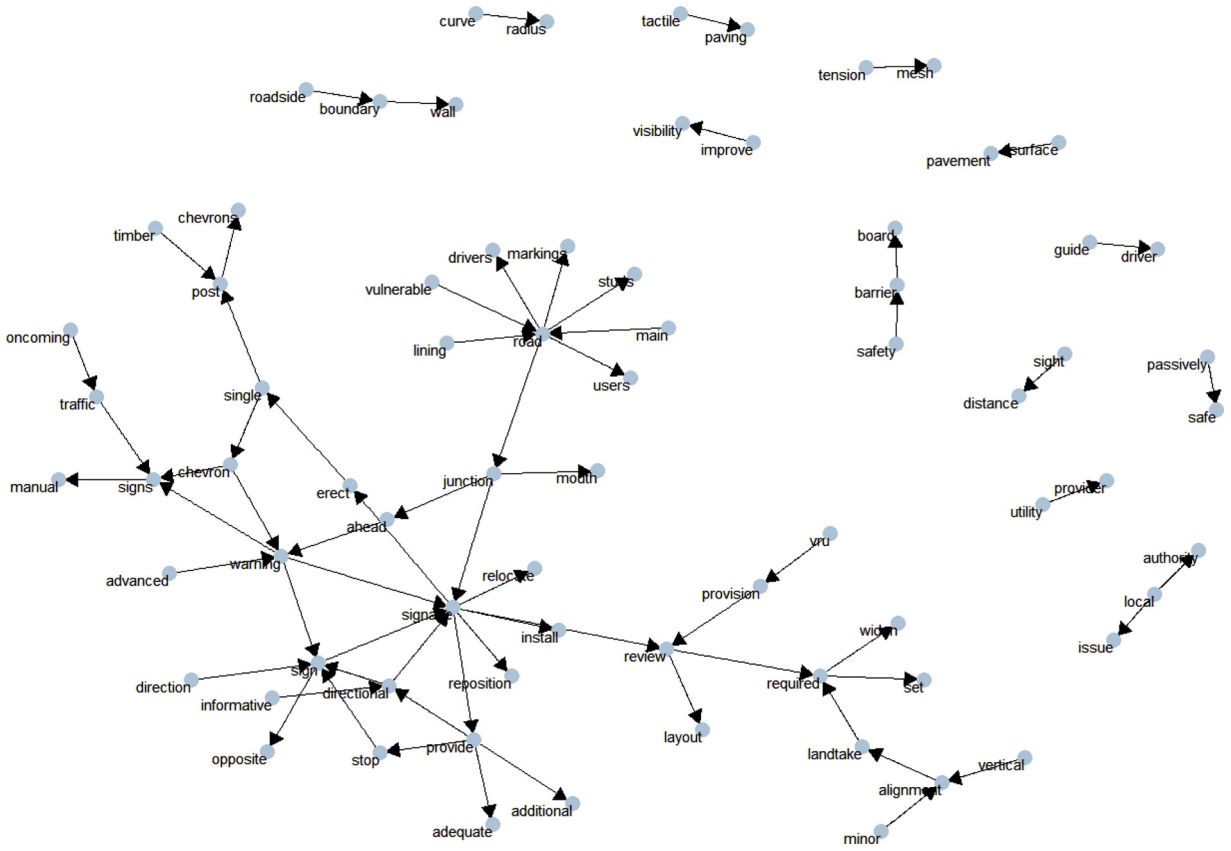

Fig. 3. Directed graph of common bigrams in the solutions record set.

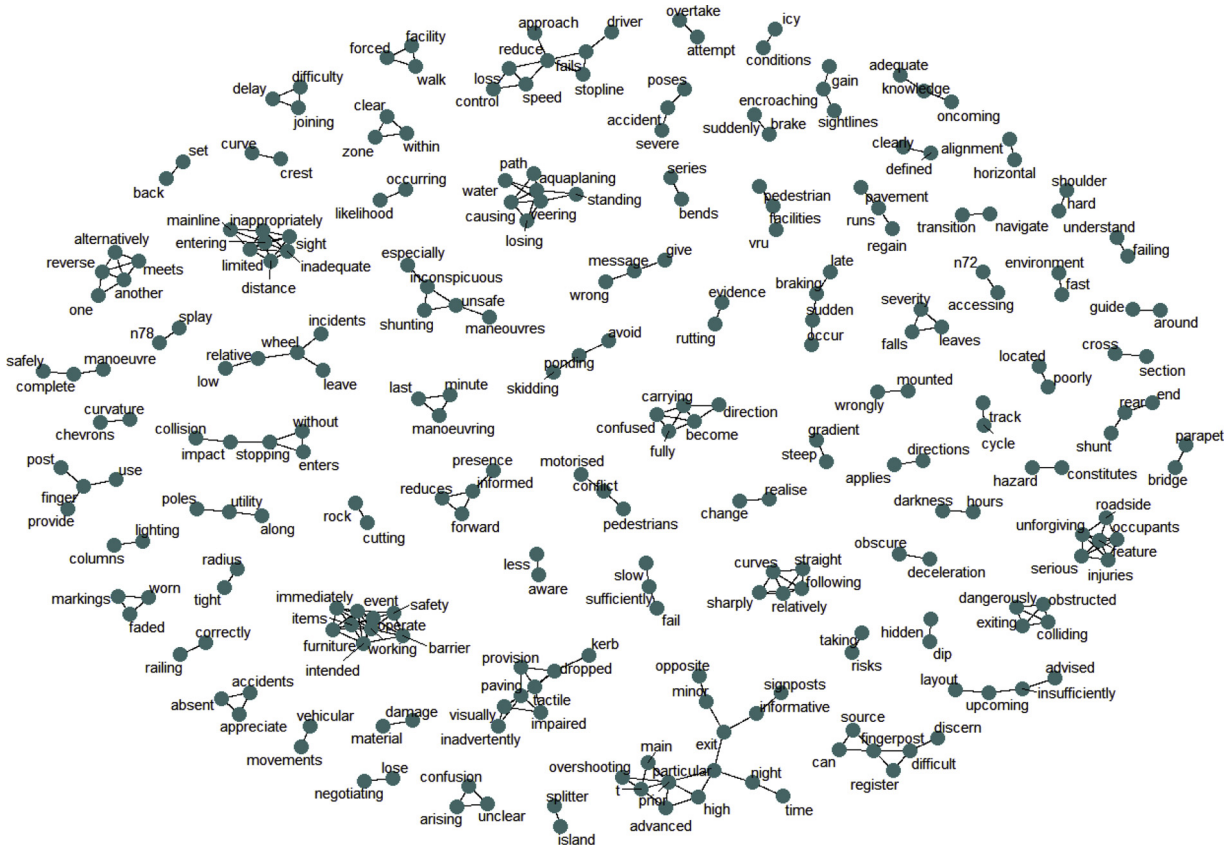

Fig. 4. Pairs of words in the problems record set that show at least a 0.50 correlation of appearing within the same document.

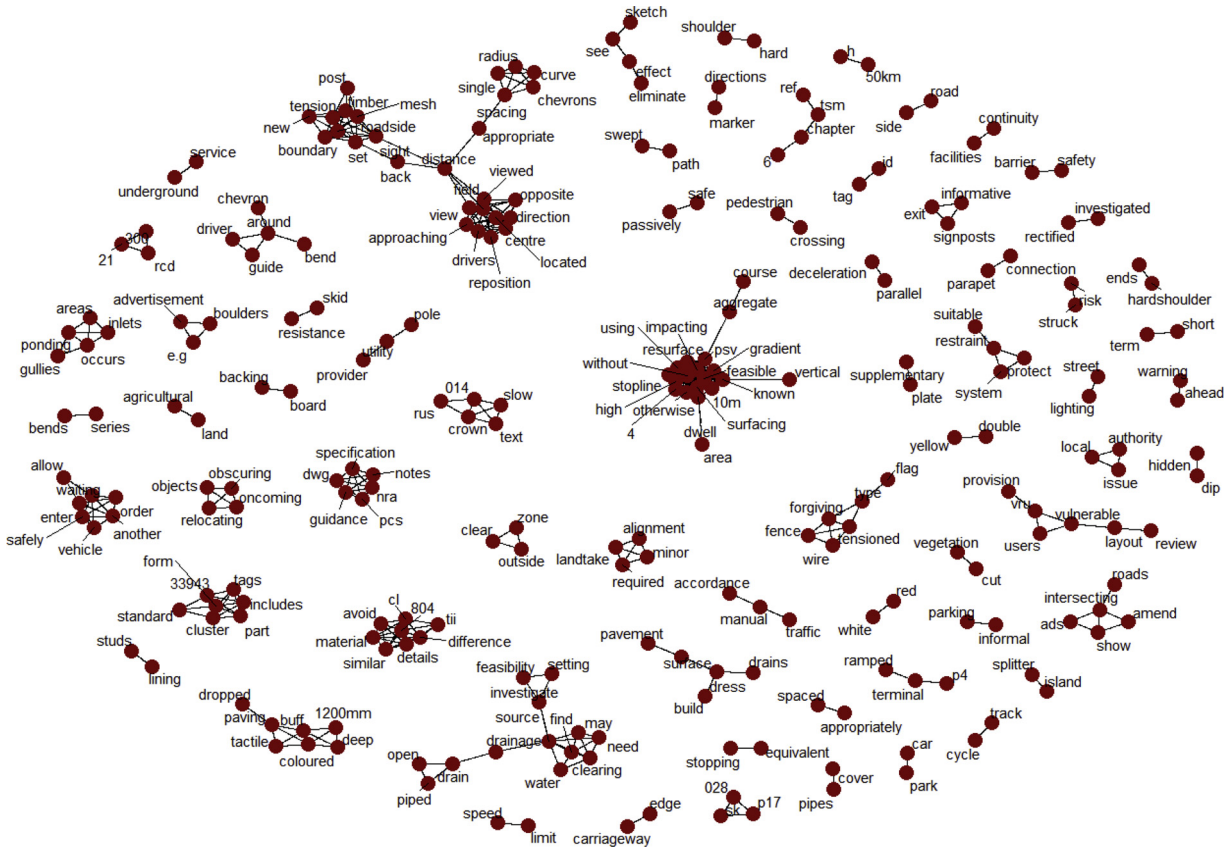

Fig. 5. Pairs of words in the solutions record set that show at least a 0.50 correlation of appearing within the same document.

Table 2 and Table 3.

To provide a better understanding of the LDA's latent topics, Fig. 8 presents some examples of the topic-specific words probabilities (β) for the 25 topics of the problems record set. For instance, the word "hazard" has a 13% probability of being generated from Topic 1, whereas "cyclists" has a 3% probability of being generated from the same topic. Fig. 9 presents the topic-specific words probabilities (β) for the 25 topics of the solutions record set. Here we can see that the word "hazard" has an 11% probability of being generated from Topic 11, whereas "poles" has a 3% probability of being generated from the same topic.

As demonstrated by Table 2 and Table 3, the extracted 25 topics obtained from both record sets match the typical issues and interventions in road safety reasonably well, suggesting that RSI reports have been successfully covering most of the relevant state-of-the-practice road safety aspects.

There are four topics in Table 2 and Fig. 8 directly related to

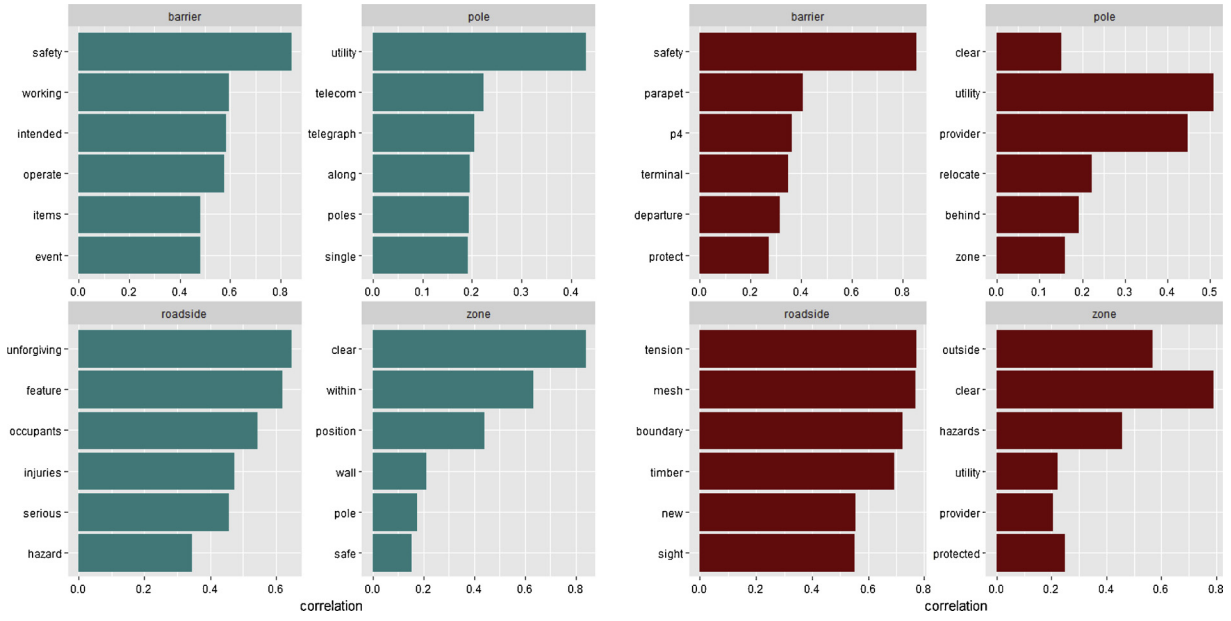

Fig. 6. Words most associated with "barrier", "pole", "roadside" and "zone" in (left) the problems record set and (right) the solutions record set.

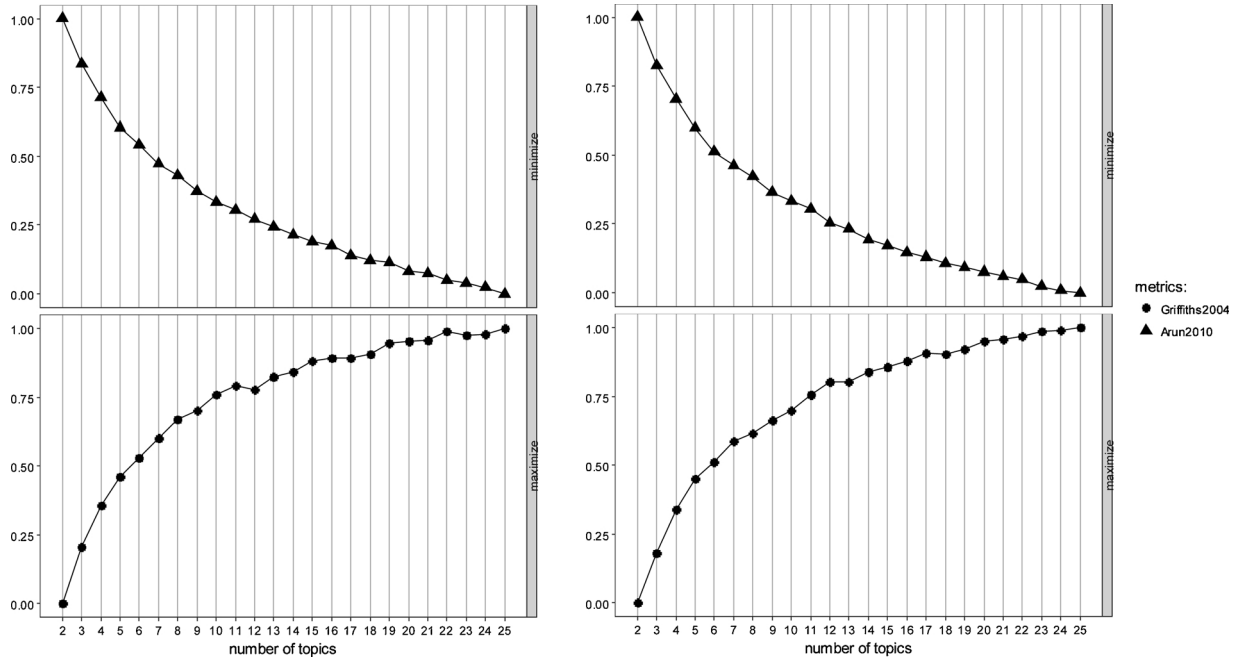

Fig. 7. Determining the number of latent topics (K) for the problems record set (left) and the solutions record set (right).

roadside problems:

- Topic 1, relates to roadside hazards and verges – on the edge of carriageway;
- Topic 2, corresponds to ROR crashes in bridges;
- Topic 5, is associated with a specific roadside hazard (poles);
- Topic 20 shows patterns of fixed continuous roadside hazards (walls and fences).

There are three topics in Table 3 and Fig. 9 directly related to roadside interventions:

- Topic 11, corresponding to mitigating the effects of roadside hazards and poles;
- Topic 14, addressing fixed roadside hazards;
- Topic 23, referring to roadside issues in bridges.

It should be noted that some words relating to the “forgiving roadside” and “clear zone” concepts, as well as the relevant European technical standards (CEN-EN1317 and EN 12,767), are absent from these topics. That is, these words are not mentioned often enough to be extracted as a distinctive topic. The absence of “clear” and “zone” in these topics may reveal a lack of application of this concept on Irish roads, at least as a specific characteristic deserving to be explicitly mentioned in the reports. Alternatively, the concept may be referred to using different words (though this is not evidenced in Figs. 4, 5, and 6), or there can even be misclassification by human coders. RSI reports were human-coded and thus might contain errors. In the latter, it is difficult to differentiate between the origins of misclassification, i.e., it is not sure whether a problem or solution is misclassified for technical or conceptual reasons.

It is also worthwhile to point out that the word “cyclists”, not typically related to roadside safety, appears in Topic 1 of the problems record set. In fact, several statements in the problems record set relate cyclists to roadside safety. Some examples are presented below:

- "Safety Barrier. Barrier layout. The pedestrian guardrails are located too close the carriageway edge resulting in insufficient lateral clearance and poses a hazard to cyclists."
- "VRU Pedestrian facilities. The location of the gantry signal pole may cause an obstruction to cyclists using this path."
- "Roadside hazard. Lighting Columns. The lamp column is very close to the edge of the pavement which may pose a hazard to vehicles, including cyclists, when meeting oncoming traffic at this location."
- "Roadside hazard. Sign Supports. The traffic lights ahead sign is located very close to the edge of the pavement. This poses a hazard to errant vehicles and cyclists."
- "Roadside hazard. Utility poles along carriageway. The telephone pole is located very close to the edge of the carriageway. Acts as a potential hazard to errant vehicles and cyclists."
- "Roadside hazard. Boundary walls. The retaining wall and pedestrian railings are located very close to the edge of the carriageway. This poses a hazard to errant vehicles and cyclists."
- "Utility poles along carriageway. The lighting column in the carriageway constitutes a hazard to vehicles including cyclists".

The examples shown indicate that vulnerable road users’ safety issues are already considered in Irish RSI, even though issues for vehicle occupants and drivers are still more frequently mentioned.

While LDA models estimate each topic as a combination of words (with probabilities of β), it also estimates each Document as a combination of topics (with probabilities of γ).

To identify the frequency distribution of the topics in both *corpus* (problems and solutions record sets), a heuristic assumption was introduced where each Document should be categorized into one, and only one, topic group. That is, each Document is categorized into the topic group that shows the highest γ value. With this assumption, the relative frequencies of the topics are shown in Table 4. The shaded cells represent the roadside safety topics identified in Table 2 and Table 3 as most relevant for roadside safety. From Table 4, it is clear that the frequency of topics related to roadside safety is higher in the problems record set (topics 1, 2, 5 and 20 which correspond to 20%) than in the solutions record set (topics 11, 14 and 23 which correspond to 17%), meaning that problems are more easily identified and related to the roadside area than interventions may be. This seems reasonable, as sometimes roadside safety issues may be mitigated by interventions on the mainline of the roadway itself (e.g. improving road surface characteristics and correcting geometric deficiencies). The relative

Table 2
Extracted Latent Topics with keywords (problems record set).

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	hazard	vehicle	collision	road	barrier	sign	may	sign	through	pedestrians
2	edge	errant	side	layout	vehicle	junction	layout	may	road	pedestrian
3	carriageway	parapet	mainline	users	safety	warning	lack	result	side	conflict
4	constitutes	bridge	impact	inappropriate	pole	located	drivers	missing	traffic	crossing
5	Roadside-hazard	unprotected	without	resulting	along	ahead	definition	being	onto	between
6	cyclists	strike	where	type	poles	advance	upcoming	warning	potential	facilities
7	hard	striking	stopping	clearly	errant	poorly	which	ahead	layout	carriageway
8	over	increased	enters	potential	within	too	result	been	see	motorised
9	area	severity	vehicle	other	lighting	misleading	aware	provided	conflicts	impaired
10	verge	drop	risk	defined	has	visible	being	has	result	footway
	Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
1	vehicles	due	may	road	loss	bend	risk	end	mainline	vehicles
2	left	junction	drivers	collisions	control	control	road	rear	entering	risk
3	exiting	drivers	signs	type	speed	loss	due	shunt	drivers	boundary
4	visibility	main	roundabout	signage	approach	risk	poor	from	visibility	wall
5	right	approaching	lead	failing	where	speed	control	risk	inadequate	roadside
6	lane	night	insufficient	leading	driver	chevrons	surface	directional	sight	collisions
7	from	particular	post	night	fails	lose	water	minor	distance	fence
8	turn	high	finger	chevron	reduce	does	ponding	braking	which	occupants
9	turning	exit	information	time	stopline	change	oncoming	shunts	result	injuries
10	junction	risk	which	both	risk	negotiating	avoid	exit	alignment	errant

	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
1	junction	road	driver	carriageway	risk
2	braking	stop	risk	could	side
3	turning	side	from	type	approaching
4	sudden	traffic	approach	head	collision
5	unsafe	markings	confusion	very	where
6	late	line	location	lead	driver
7	direction	overtaking	side	vehicle	vehicles
8	manoeuvres	steep	late	section	impact
9	difficult	gradient	braking	cross	see
10	shunting	centre	road	potential	fails

frequencies per solutions topic are higher than per problems topic, reflecting the lower variety of effective tools that road inspectors may select to address detected problems.

5. Conclusions

In this paper, a framework to identify many-to-many associations among a broad group of conditions associated with ROR crashes and roadside safety interventions in plain text standardized road safety inspection reports is proposed. The text mining technique was used to

Table 3
Extracted Latent Topics with keywords (solutions record set).

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	junction	junction	sign	road	may	risk	driver	end	pedestrian	bend
2	warning	signs	located	layout	result	control	from	rear	pedestrians	control
3	sign	unsafe	stop	type	being	loss	risk	shunt	carriageway	speed
4	ahead	turning	signinglining	resulting	upcoming	where	approach	risk	crossing	risk
5	advance	directional	poorly	inappropriate	layout	over	side	due	facilities	loss
6	overshooting	direction	too	potential	which	approach	braking	turning	conflict	drivers
7	local	late	junction	may	lack	pavement	location	delay	lack	approaching
8	having	post	misleading	users	definition	strikes	confusion	collisions	may	poor
9	brake	difficult	obscured	clearly	missing	runs	late	from	footway	due
10	does	shunting	visible	defined	insufficiently	rock	road	braking	impaired	chevrons
	Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
1	carriageway	road	side	wall	vehicles	mainline	collision	road	speed	could
2	hazard	due	approaching	boundary	visibility	entering	side	through	driver	lead
3	edge	main	risk	errant	exiting	drivers	mainline	side	approach	head
4	constitutes	exit	collision	roadside	left	sight	where	onto	fails	very
5	vehicle	night	impact	vehicle	alignment	inadequate	vehicle	see	reduce	width
6	poles	drivers	driver	within	from	distance	without	traffic	where	section
7	leave	particular	where	risk	crest	which	impact	potential	loss	struck
8	pole	high	see	occupants	forward	result	risk	conflicts	control	area
9	roadsidehazard	minor	fails	increased	right	visibility	stopping	layout	stopline	cross
10	along	risk	vehicles	injuries	mainline	limited	enters	possible	does	narrow

	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
1	may	road	vehicle	vehicles	road
2	drivers	poor	errant	between	collisions
3	sudden	markings	barrier	left	type
4	roundabout	both	parapet	conflict	failing
5	overshoot	line	bridge	lane	signage
6	occur	lighting	safety	traffic	leading
7	braking	risk	strike	turn	chevron
8	manoeuvres	water	severity	motorised	time
9	road	surface	striking	insufficient	ahead
10	unaware	directions	increased	cyclists	layout

analyse 54 RSI reports, covering 4011 km of roads, and gathered over six years (2012–2017). The combination of text mining and topic modelling enables the RSI report problems and solutions to be extracted and evaluated. LDA was applied to two datasets collected from the reports, one comprising the issues raised, and the other the proposed solutions. The results of this study provide evidence that the LDA is methodologically appropriate for identifying the co-occurrence patterns of attributes related to the ROR crashes in road safety inspections' reports, as well as the interventions' patterns associated with these crashes.

Our results showed that important keywords relating to the "forgiving roadside" and "clear zone" concepts, as well as the relevant European technical standards (CEN-EN1317 and EN 12,767) are absent from the extracted latent topics. That is, these words are not mentioned

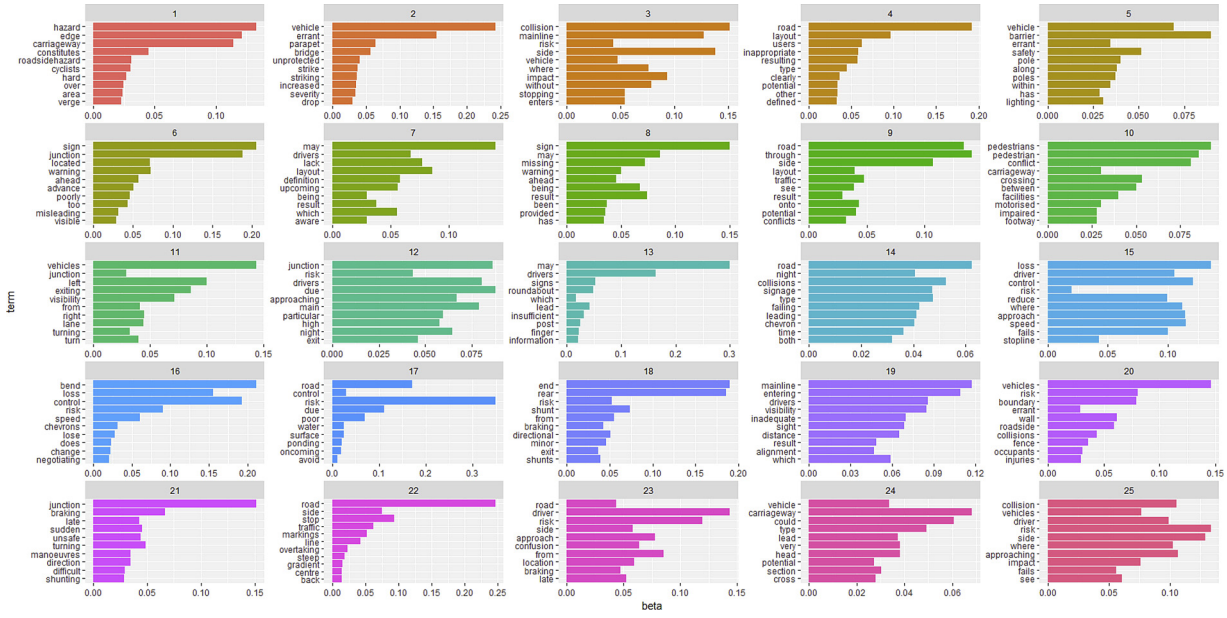

Fig. 8. Topic-specific word probabilities for the problems record set.

often enough to be extracted as a distinctive topic. The absence of the "clear zone" term in those topics may reveal a lack of application of this concept in Irish RSI reports, at least as a desirable specific characteristic deserving to be explicitly mentioned by a technical term in the reports.

It is also worthwhile to point out that the word "cyclists", which is not typically related to roadside safety on high speed roads, appears in one of the topics of the problems record set. In fact, several statements in that record set relate cyclists to roadside safety. Vulnerable road users' safety issues are already considered in Irish RSI, even though they are still mostly associated with issues specific to vehicle occupants and drivers.

The frequency of topics related to roadside safety is higher in the problems record set than in the solutions record set, meaning that problems are more easily identified and related to the roadside area than interventions may be. This seems reasonable, as sometimes roadside safety issues may be mitigated by interventions in the roadway itself (e.g. improving road surface characteristics and correcting geometric deficiencies).

The results of these analyses aimed to help to determine the extent to which national road authorities in Europe and their contractors are currently capable of implementing and maintaining compliance with roadside standards and guidelines throughout the life cycle of roads.

The findings of this study may also encourage policy makers to make the necessary effort to develop recommendations for safe roadside design and management.

The results of this study confirm the potential of LDA in analysing roadside safety issues. Yet this study has some limitations that need to be addressed in future studies.

First, we only focused on Irish RSI, which might have excluded specific roadside safety issues from other European countries. Therefore, caution should be exercised when generalizing some of the study conclusions to different ROR crash scenarios. The investigation of a more significant number of RSI reports and particularly from a wider range of countries will be required in future research, to the enable

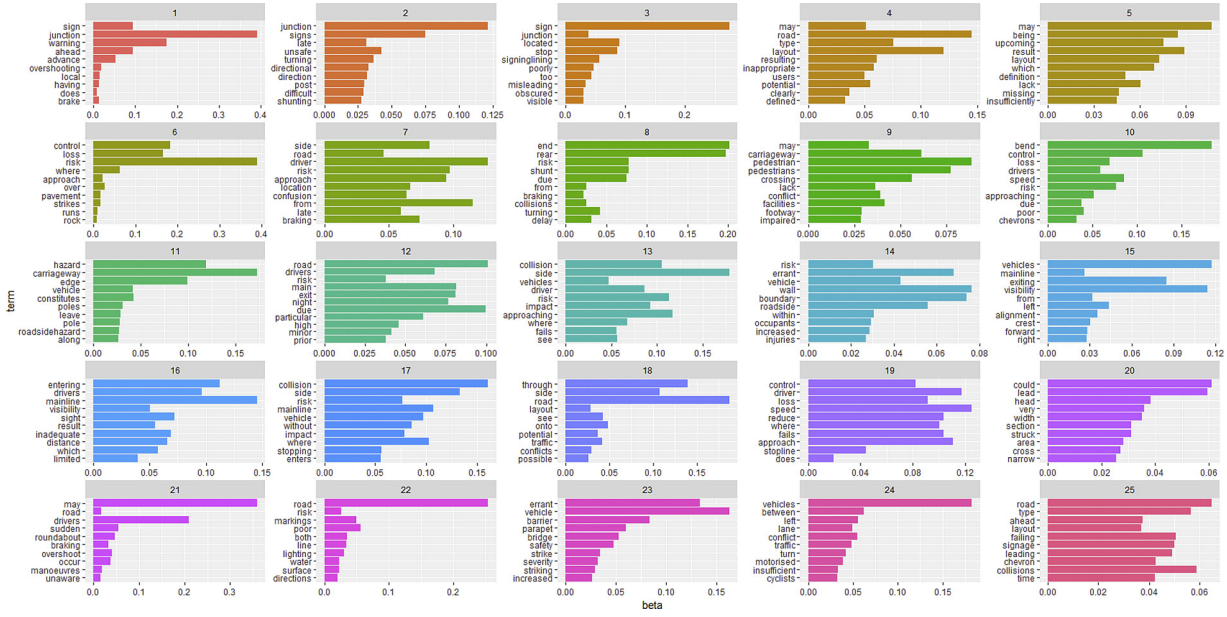

Fig. 9. Topic-specific word probabilities for the solutions record set.

Table 4
Relative frequencies of extracted topics for problems and solutions record sets.

Topic	Problems		Solutions	
	#	%	#	%
1	533	5,5%	341	3,6%
2	687	7,2%	401	4,2%
3	556	5,8%	463	4,8%
4	358	3,7%	275	2,9%
5	421	4,4%	326	3,4%
6	409	4,3%	229	2,4%
7	321	3,3%	223	2,3%
8	300	3,1%	453	4,7%
9	262	2,7%	507	5,3%
10	667	6,9%	341	3,6%
11	383	4,0%	567	5,9%
12	474	4,9%	519	5,4%
13	179	1,9%	446	4,6%
14	290	3,0%	401	4,2%
15	635	6,6%	354	3,7%
16	355	3,7%	387	4,0%
17	127	1,3%	374	3,9%
18	411	4,3%	241	2,5%
19	443	4,6%	639	6,7%
20	287	3,0%	332	3,5%
21	431	4,5%	298	3,1%
22	270	2,8%	285	3,0%
23	191	2,0%	625	6,5%
24	252	2,6%	363	3,8%
25	364	3,8%	215	2,2%
Total	**9606**	**100%**	**9605**	**100%**

expansion of the models and the generalization of their interpretation. Secondly, RSI reports from other road categories could be included in the analysis (both in aggregated and disaggregated samples). Recent studies revealed problems in current roadside design, especially with regard to clear zones criteria, forgiving slopes and warrants for safety barrier installation (Roque et al., 2015; Roque and Jalayer, 2018). Future studies might consider more reports regarding dual carriageway roads and, specifically, motorways (only 0.6% of our sample data) to further explore these issues and thereby increase roadside safety in the most efficient way.

Possible extensions of this study may focus on the development of similar analyses based on data collected for other types of crashes, such as multi-vehicle crashes. Additionally, surveys may involve National Road Authorities and practitioners working in roadside safety management and design, to collect their experiences and to assess if their opinions are in accordance with the conclusions from this study.

For future work, it would also be interesting to apply models that have the potential to link the topics with specific roadside problems or that can avoid the early establishment of the number of topics. Traditional unsupervised topic models such as LDA are intended only to use the discrete bag-of-words representation and cannot explore metadata that might be available for each RSI report (Kandemir et al., 2018). In RSI reports, metadata can represent star ratings for pedestrians, cyclists, motorcyclists and vehicle occupants, or document timestamps. Thus, other modelling approaches can be explored, such as, structural topic models (Roberts et al., 2014, 2016), which allow topics to correlate when estimating the model, and graph topic models (Xuan et al., 2015; Zhang et al., 2019) that releases the assumption of 'bag of words' and assume a graph structure for the text. Also, there is merit in carrying out similar analyses using Bayesian nonparametric topic models (Blei et al., 2010; Williamson et al., 2010), which can not only learn the summarised topics in a set of documents but also adapt the number of learned topics according to the documents in the set (Xuan et al., 2019).

Declaration of Competing Interest

None.

Acknowledgements

The authors wish to thank the PROGReSS Project (PROvision of Guidelines for Road Side Safety) funded under the CEDR Transnational Road Research Programme 2016, by the national directors of roads in Austria, Belgium (Flanders), Finland, Ireland, Netherlands, Norway, Slovenia, Sweden and the United Kingdom.

Carlos Roque gratefully acknowledges the scholarship SFRH/BPD/118499/2016 by the Portuguese Science and Technology Foundation Agency (FCT - *Fundação para a Ciência e a Tecnologia, IP*).

The authors also express their gratitude to the three anonymous reviewers, for their valuable suggestions that have helped to improve this paper to its final version.

References

Arnon, I., Snider, N., 2010. More than words: frequency effects for multi-word phrases. J. Mem. Lang. 62 (1), 67–82.
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. Springer, Berlin, Heidelberg, pp. 391–402.
Bao, J., Liu, P., Qin, X., Zhou, H., 2018. Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data. Accid. Anal. Prev. 120, 281–294.
Bastani, K., Namavari, H., Shaffer, J., 2019. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Syst. Appl. 127, 256–271.
Bhattacharya, M., Jurkovitz, C., Shatkay, H., 2018. Co-occurrence of medical conditions: exposing patterns through probabilistic topic modeling of snomed codes. J. Biomed. Inform. 82, 31–40.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Blei, D.M., Griffiths, T.L., Jordan, M.I., 2010. The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies. Journal of the ACM (JACM) 57 (2), 7 (2010).
Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S., 2009. A density-based method for adaptive LDA model selection. Neurocomputing 72 (7–9), 1775–1781.
Canito, J., Ramos, P., Moro, S., Rita, P., 2018. Unfolding the relations between companies and technologies under the Big Data umbrella. Comput. Ind. 99, 1–8.
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.
Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C., Blei, D.M., 2009. Reading tea leaves: how humans interpret topic models. Advanced Neural Information Processing Systems 288–296.
Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41 (6), 391–407.
Deveaud, R., SanJuan, E., Bellot, P., 2014. Accurate and effective latent concept modelling for ad hoc information retrieval. Document Numérique 17 (1), 61–84.
DiMaggio, P., Nag, M., Blei, D., 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. 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. Journal of Accounting and Economics 64 (2-3), 221–245.
Eenink, R., Reurings, M., Elvik, R., Cardoso, J.L., Wichert, S., Stefan, C., 2007. Accident Prediction Models and Road Safety Impact Assessment: Recommendations for Using These Tools. Deliverable D2 Leidschendam, The Netherlands.
European Union, 2008. Strasbourg, FranceDirective 2008/96/EC of the European Parliament and of the Council of 19th November 2008 on Road Infrastructure Safety Management2008. Directive 2008/96/EC of the European Parliament and of the Council of 19th November 2008 on Road Infrastructure Safety Management.
Feinerer, I., Hornik, K., Meyer, D., 2008. Text mining infrastructure in R. J. Stat. Softw. 25 (5), 1–54.
Ghosh, D., Guha, R., 2013. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. Cartogr. Geogr. Inf. Sci. 40 (2), 90–102.
Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. 101, 5228–5235.
Grun, B., Hornik, K., 2011. Topic models: an R package for fitting topic models. J. Stat. Softw. 40 (13), 1–30.
Hamed, T., Dara, R., Kremer, S.C., 2018. Network intrusion detection system based on recursive feature addition and bigram technique. Comput. Secur. 73, 137–155.
Hauer, E., 1998. Knowledge and the management of safety. Traffic Safety Summit.
Hofmann, T., 1999. Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval 50–57.
Ihou, K.E., Bouguila, N., 2019. Variational-based latent generalized Dirichlet allocation

model in the collapsed space and applications. Neurocomputing 332, 372–395.
ITF, 2016. Zero Road Deaths and Serious Injuries, Leading a Paradigm Shift to a Safe System. International Transport Forum. Éditions OCDE, Paris, France.
Kandemir, M., Kekeç, T., Yeniterzi, R., 2018. Supervising topic models with Gaussian processes. Pattern Recognit. 77, 226–236.
Li, R., Pereira, F.C., Ben-Akiva, M.E., 2015. Competing risk mixture model and text analysis for sequential incident duration prediction. Transport. Res. Part C: Emerg. Technol. 54 (2015), 74–85.
McAuliffe, J.D., Blei, D.M., 2008. Supervised topic models. Advances in Neural Information Processing Systems. pp. 121–128.
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.
Matena, S., Weber, R., Huber, C., Hruby, Z., Pokorny, P., Gaitanidou, E., Vaneerdewegh, P., Strnad, B., Cardoso, J.L., Schermers, G., Elvik, R., 2007. Road Safety Audit – Best Practice Guidelines. Qualification for Auditors and Programming. Ripcord-Iserest Report D4.2. BASt.
Nikita, M., 2016. Tuning of the Latent Dirichlet Allocation Models Parameters. R Package Ldatuning Version 0.2.0. Comprehensive R Archive Network (CRAN).
Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2013. Text analysis in incident duration prediction. Transp. Res. Part C Emerg. Technol. 37, 177–192.
Proctor, S., Belcher, M., Cook, P., 2003. Practical Road Safety Auditing. Thomas Telford, London ISBN: 0727729381.
Qi, G., Guan, W., 2019. Quantitatively mining and distinguishing situational discomfort grading patterns of drivers from car-following data. Accid. Anal. Prev. 123, 282–290.
R Development Core Team, 2011. R: a Language and Environment for Statistical Computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D., 2014. Structural topic models for open-ended survey responses. Am. J. Pol. Sci. 58 (4), 1064–1082.
Roberts, M.E., Stewart, B.M., Airoldi, E.M., 2016. A model of text for experimentation in the social sciences. J. Am. Stat. Assoc. 111 (515), 988–1003.
Robinson, S.D., 2019. Temporal topic modeling applied to aviation safety reports: a subject matter expert review. Saf. Sci. 116, 275–286.
Roque, C., Jalayer, M., 2018. Improving roadside design policies for safety enhancement using hazard-based duration modeling. Accid. Anal. Prev. 120, 165–173.
Roque, C., Moura, F., Cardoso, J.L., 2015. Detecting unforgiving roadside contributors through the severity analysis of ran-off-road crashes. Accid. Anal. Prev. 80, 262–273.
Salton, G., McGill, M., 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY.
Selby, P., Earhart, V., Garrison, E., Douglas Raymer, G., 2004. Tests of induction in mice by acute and chronic ionizing radiation and ethylnitrosourea of dominant mutations that cause the more common skeletal anomalies. Mutat. Res. Mol. Mech. Mutagen. 545 (1–2), 81–107.
Silge, J., Robinson, D., 2017. Text Mining With R - a Tidy Approach. O'Reilly Media, Inc., Sebastopol, CA.
Sørensen, M., Elvik, R., 2007. Best Practice Guidelines on Black Spot Management and Safety Analysis of Road Networks. RIPCORD-ISEREST Deliverable D6, Oslo, Norway.
Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical dirichlet processes. J. Am. Stat. Assoc. 101, 1566–1581.
Transport Infrastructure Ireland, 2014. NRA HA 17 Road safety inspection guidelines. AM-STY-06043 Road Safety Inspection Guidelines. TII Publications.
Wang, W., Feng, Y., Dai, W., 2018. Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. Electron. Commer. Res. Appl. 29, 142–156.
Williamson, S., Wang, C., Heller, K.A., Blei, D.M., 2010. The IBP compound dirichlet process and its application to focused topic modeling. Proceedings of the 27th International Conference on Machine Learning (ICML-10).
Xuan, J., Lu, J., Zhang, G., Luo, X., 2015. Topic model for graph mining. IEEE Trans. Cybern. 45 (12), 2792–2803.
Xuan, J., Lu, J., Zhang, G., 2019. A survey on bayesian nonparametric learning. ACM Computing Surveys (CSUR) 52 (1), 13 2019.
Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. Transp. Res. Part C Emerg. Technol. 86, 580–596.
Zhang, H., Huating, S., Wu, X., 2019. Topic model for graph mining based on hierarchical Dirichlet process. Stat. Theory Relat. Fields. https://doi.org/10.1080/24754269.2019.1593098.
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. 16 (13), S8.