8033 words in the main text
271 words in the abstract
70 references
3 tables and 8 figures in the main text

Text as data: narrative mining of non-collision injury incidents on public buses by structural topic modeling

Pengpeng Xu[a,b], Qianfang Wang[a], Yun Ye[c,d,e], S.C. Wong[f], Hanchu Zhou[g†]

a School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China
b Hunan Key Laboratory of Smart Roadway and Cooperative Vehicle-Infrastructure Systems, Changsha University of Science & Technology, Changsha, China
c Faculty of Maritime and Transportation, Ningbo University, Ningbo, China
d Collaborative Innovation Center of Modern Urban Traffic Technologies, Southeast University, Nanjing, China
e National Traffic Management Engineering & Technology Research Center Ningbo University Sub-Center, Ningbo, China
f Department of Civil Engineering, The University of Hong Kong, Hong Kong, China
g School of Traffic and Transportation Engineering, Central South University, Changsha, China
†Corresponding author; Email: hanchuzhou@csu.edu.cn

Funding: This work was supported by grants from the National Natural Science Foundation of China (Project No. 52302433), Natural Science Foundation of Guangdong Province, China (Project No. 2023A1515012404; 2024A1515011578), Talent Start-up Project of Ningbo University (Project No. ZX2023000249), Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 17202824), Open Fund of Hunan Key Laboratory of Smart Roadway and Cooperative Vehicle-Infrastructure Systems, Changsha University of Science & Technology (Project No. kfj230801), and National "111" Centre on Safety and Intelligent Operation of Sea Bridges (Project No. D21013). Prof. S.C. Wong was supported by the Francis S Y Bong Professorship in Engineering. The funders had no role in the study design, data collection and processing, preparation of the manuscript, or decision to publish.

Text as data: narrative mining of non-collision injury incidents on public buses by structural topic modeling

Abstract

Introduction: Although numerous studies have investigated collisions involving public buses, there has been inadequate research on passenger injuries caused by non-collision incidents on public buses. One major obstacle is that the manual extraction of thematic information from massive document repositories is exceedingly labor intensive, cumbersome, and inaccurate. Our study thereby illustrated how to automatically characterize non-collision injury incidents on public buses by fusing advanced language processing techniques and large-scale incident reports.

Methods: Based on the 12,823 textural narratives recorded by police during 2010–2019 in Hong Kong, the structural topic modeling was developed to uncover underlying themes, quantify topic prevalence, and portray complex interconnectedness.

Results: Thirty-three topics were successfully labeled, with the topic *stand and lost balance* being the most prevalent. Non-collisions were more likely to result in serious consequences when incidents occurred because the bus skidded, when a passenger was boarding, and when a standing passenger lost the balance. Six unique patterns were uncovered, i.e., the failure to hold handrails accompanied by inappropriate behaviors of bus drivers when approaching bus stations, loss of balance among standing passengers due to the sharp braking of bus drivers in response to red traffic lights ahead, alighting passengers being hit by the door, passengers falling while climbing staircases, passengers being injured because of bus driver's emergency maneuvers to avoid collisions with nearside pedestrians, and passengers being injured due to the careless lane-changing of bus drivers when weaving through roundabouts.

Conclusions: By leveraging the emerging text mining techniques, unstructured narratives written by the police can provide valuable and organized information for regular injury surveillance. Tailor-made countermeasures were proposed to prevent non-collision injury incidents on public buses.

Keywords: Text mining; structural topic modeling; network topology; non-collision injuries; public bus

1. Introduction

Many cities worldwide encourage the use of public buses to promote sustainable, accessible, and equitable mobility in urban settings (Tao et al., 2024). In Hong Kong, approximately 30% of commuting trips are made via public buses (HKTD, 2014), and more than 1800 passengers are unexpectedly injured on public buses each year (HKTD, 2024). Although the share of traffic injuries involving public bus passengers is low, the risk of injury on a bus is far from negligible (Akintayo and Adibeli, 2022). Previous studies on public bus safety have focused primarily on collisions with roadside objects, pedestrians, cyclists, or other vehicles, as these incidents are expected to result in more serious consequences (Chen et al., 2022). Recently, researchers have shifted their attention to non-collision injuries on public buses because of their overrepresentation (Kendrick et al., 2015; Elvik, 2019; Chen et al., 2024). For example, in Hong Kong, nearly 70% of injuries to public bus passengers are attributed to non-collision incidents (Zhou et al., 2020), whereas in Tanzania, 71% of bus commuters have experienced non-collision injuries during their lifetime (Lwanga et al., 2022). These non-collision incidents typically occur when onboard passengers fall due to improper maneuvers by bus drivers such as sudden decelerations or sharp turning, and when passengers fall while boarding or alighting from the bus (Elvik, 2019).

Over the past decades, a few studies have investigated the determinants of passenger falls on public buses by conducting biomechanical experiments under controllable, reproducible, and ideal conditions (Karekla and Tyler, 2018; 2019; Arabian et al., 2020; Karekla and Fang, 2021; Elawad et al., 2024). Most research, however, has relied on available injury records routinely reported by local authorities such as the police (Barnes et al., 2016; Zhou et al., 2020) or hospitals (Björnstig et al., 2005; Halpern et al., 2005; Zunjic et al., 2012; Silvano and Ohlin, 2019; Siman-Tov et al., 2019; Chen et al., 2024). Through a descriptive analysis of established variables captured using forced-choice fields such as checkboxes and dropdown menus, contextual characteristics of non-collision injury incidents on public buses can be coarsely profiled. Such practice, however, may be inadequate, because the injury mechanism is unlikely to be completely captured by the restrictive, predefined options of tabular forms (Lopez et al., 2022). This is particularly true for non-collision injury incidents compiled by the police, given that the reporting system is designated specifically for the collection of collision incidents (Abay, 2015). In comparison, open-ended narratives, which have long been an integral part of injury report templates, may provide valuable yet hidden information in addition to the organized tabulated data (Ahmed et al., 2019; Alambeigi et al., 2020; Arteaga et al., 2020; Kwayu et al., 2021; Wali et al., 2021; Goldberg, 2022; Kutela et al., 2022a; Wang et al., 2022), even though these narratives have not been fully utilized. One major obstacle might be that the manual examination of unstructured text narratives and extraction of thematic information from massive document repositories one by one are exceedingly labor intensive, cumbersome, and inaccurate.

Fortunately, with the unprecedented breakthroughs in natural language processing techniques that facilitate the automatic identification of categories or latent themes within large document corpora (Grimmer and Stewart, 2013; Cambria and White, 2014; Young et al., 2018), descriptive narratives written by the police might be revitalized as a complementary source of information. As the police has recorded in detail what occurred before, during, and after an incident by answering *who*, *what*, *why*, and *how* questions, these narratives have a huge potential to characterize the nature of non-collision injury incidents on public buses by leveraging emerging text-mining techniques such as the structural topic modeling (STM; Roberts et al., 2016; Roberts et al., 2019).

Unlike manual labeling that depends strongly on a subjective and arbitrary predefinition of association rules, topic modeling can uncover latent topics across all documents more efficiently and objectively by identifying clusters of words. Recent studies have extensively demonstrated topic modeling as a powerful and promising tool for automatic mining of versatile textual data, such as white papers (Bongini et al., 2022), newspapers (Adämmer and Schüssler, 2020; Huang and Loo, 2023), open-ended survey responses (Roberts et al., 2014; Baburajan et al., 2022; Kutela et al., 2022b), medical prescriptions (Zafari and Ekin, 2019), Facebook feeds (Ravenda et al., 2022), Twitter feeds (Ramondt et al., 2022), Sina Weibo feeds (Jing et al., 2023), aviation incident reports (Kuhn, 2018; Rose et al., 2022), and transit card data (Aminpour and Saidi, 2025). In road safety domain, pioneering research has also been conducted to garner in-depth insights into crash causes by mining official reports (Alambeigi et al., 2020; Kwayu et al., 2021; Wali et al., 2021). Specifically, Alambeigi et al. (2020) used probabilistic topic modeling to characterize crashes involving automated vehicles by mining 114 narratives released by the California Department of Motor Vehicles from October 214 to June 2019. Five topics pertaining to transitions of control, collisions at intersections in a right-turn lane, collisions at intersections in non-turn lanes, sideswipe crashes during a left-overtake, and collisions at intersections involving oncoming traffic were successfully identified. Despite being informative, their study hardly drawn definitive conclusions given a limited sample size. Wali et al. (2021) applied factor analysis to generate 13 new variables that were not captured by traditional tabular forms from 6470 crash narratives, which substantially enriched the results of injury severity analysis for pedestrian and bicyclist trespassing crashes. Similarly, based on 9209 crash narratives recorded by the police in Michigan, United States, Kwayu et al. (2021) used STM to reveal the prevalence of latent themes in fatal crash narratives. Their proposed framework extended understanding of contextual factors associated with fatal crashes. For example, topics such as involvement of passengers, crossing the centerline, driving under the influence, and speeding were prevalent for fatal crashes involving young drivers, while topics such as turning left, failure to yield, and lane changing were closely related with fatal crashes involving older drivers.

To propose effective safety strategies, decision makers need to explicitly understand the complex relationships among various contributing factors. By mapping the generated topics and connections between topics as nodes and edges, respectively, network topology analysis allows the visualization of complex interconnectedness between topics in terms of network properties and node centrality measures (Radicchi et al., 2004; Kwayu et al., 2021; Kutela et al., 2022a). As a non-collision injury incident is an aggregative consequence of multiple factors, the topological network is also capable of revealing common patterns by identifying topics that tend to co-occur frequently (Chang et al., 2019; Liu and Yang, 2022). Herein, to demonstrate how valuable and organized information can be automatically extracted from massive, unstructured, and open-ended injury reports, based on the 12,823 narratives recorded by the police within the last decade in Hong Kong, we attempt to mine the underlying themes of non-collision injury incidents on public buses, to portray their dynamic patterns, and to untangle their interactions and co-occurrences by leveraging STM and network typology techniques. To the best of knowledge, our study is among the first to uncover the unique features of non-collision injury incidents on public buses by integrating state-of-the-art natural language processing techniques and large-scale textual data. This analysis not only aids to propose a range of evidence-based, actionable countermeasures to productively curb public bus passenger injuries resulting from non-collision incidents, but also incentivizes stakeholders to rethink and reshape their current incident reporting paradigms, thereby promoting the penetration of digitized narratives for regular safety monitoring.

The remainder of the paper is organized as follows. After a detailed description of the formulated topic modeling technique, the data collected for analysis are introduced and processed. We then present and interpret results, demonstrate the potential of STM for semantic mining of unstructured and unorganized injury narratives, acknowledge the limitations, and conclude our study with a discussion on promising directions for future studies.

2. Methods

Unlike supervised learning that trains a classifier to correctly categorize a specific type of incidents (Arteaga et al., 2020; Goldberg, 2022; Liu and Yang, 2022), topic modeling, as an unsupervised learning method, enables users to identify latent topics that form a document. A topic here consists of several words, and a document is a mixture of topics. Therefore, a single document may comprise multiple topics, and topics are more likely to appear in the same document if they share similar semantically interpretable themes. In the present study, a novel topic model namely STM is harnessed to explore the pattern of incident narratives. STM synthesizes the merits of several topic modeling approaches, including the Dirichlet multinomial regression topic model, correlated topic model, and sparse additive generative topic model (Roberts et al., 2019). Another sound advantage of STM lies in its ability to uncover the relationship between topics and document metadata, which allows us to quantify the prevalence of textual topics among documents across different stratifications.

2.1 Formulation of STM

Like other topic models, as a generative model of word counts, the STM progressively generates document–topic and topic–word distributions by maximizing their likelihoods, given the metadata and prespecified functional forms. The document here is represented by a single incident narrative written by a police officer. For document d with vocabulary of size V_d, the generative process of STM with K topics is expressed as follows.

(1) A logistic-normal generalized linear model is formulated to describe the document-level probability for each topic based on a vector of words $\mathbf{X_d}$ for document d:

$$\theta'_d \left| \mathbf{X_d}\gamma, \Sigma \sim \text{Logistic Normal}\,(\mu = \mathbf{X_d}\gamma, \Sigma) \right. \quad (1)$$

where $\mathbf{X_d}$ is a $1 \times p$ vector, γ is a $p \times (K-1)$ matrix of coefficients, and Σ is a $(K-1) \times (K-1)$ covariance matrix. Compared with the latent

Dirichlet allocation adopted by Pereira et al. (2013), Hasan and Ukkusuri (2014), Roque et al. (2019), Alambeigi et al. (2020), Wang et al. (2022), Jing et al. (2023), and Aminpour and Saidi (2025) which assumes independence between topics (Blei et al., 2003), the use of a logistic-normal distribution is more flexible, as it allows topics to be correlated.

(2) Given the baseline word distribution m, a document-level covariate y_d that explains the thematic content, the topic specific deviation $k_k^{(t)}$, the covariate group deviation $k_{y_d}^{(c)}$, and the interaction between $k_k^{(t)}$ and $k_{y_d}^{(c)}$ denoted as $k_{y_d,k}^{(i)}$, the document-specific distribution over words representing topic k is then computed as:

$$\beta_{d,k} \propto \exp(m + k_k^{(t)} + k_{y_d}^{(c)} + k_{y_d,k}^{(i)}) \tag{2}$$

where m, $k_k^{(t)}$, $k_{y_d}^{(c)}$, and $k_{y_d,k}^{(i)}$ are V-length vectors with the corresponding one-on-one entry for each word in the vocabulary. In addition, the document proportions or probabilities are formulated as $\beta_{d,k} \propto \exp(m + k_k^{(t)})$ if no content covariates exist.

(3) For each word v ($v \in (1, 2, \ldots, V_d)$) in document d, we use the document-specific distribution across topics to draw the word's topic assignment:

$$z_{d,v} \propto \text{Multinomial}(\theta_d) \tag{3}$$

(4) Finally, conditional on the chosen topic $z_{d,v}$, the representative word for topic k is obtained as:

$$\beta_{k,v} \big| z_{d,v}, \beta_{d,k=z_{d,v}} \sim \text{Multinomial}(\beta_{d,k=z_{d,v}}) \tag{4}$$

2.2 Selection of the optimal number of topics

One practical challenge faced by STM is the absence of unified and incontrovertible criteria to judge the appropriate number of topics (Roberts et al., 2019; Kwayu et al., 2021; Rose et al., 2022). The optimal number of topics is typically determined by combining the judgments of domain experts and estimation results (Roberts et al., 2016; Zafari and Ekin, 2019; Kwayu et al., 2021). Fortunately, several data-driven diagnostic indicators, such as the residuals, semantic coherence, and exclusivity, can help justify the optimal number. *Residuals* measure the multinomial variance during the data generation process of STM. A model with a lower residual value is more desirable, because higher residuals suggest that more topics are required to capture extra variance present in the data. *Semantic coherence* is closely related to the pointwise mutual information, whose value is maximized when the most probable words in a specific topic frequently appear together (Mimno et al., 2011). Formally, let $D(v_i, v_j)$ be the number of times that word v_i and word v_j co-occur in a document. For topic k with the M most probable words, the semantic coherence C_k is calculated as:

$$C_k = \sum_{i=2}^{M}\sum_{j=1}^{i-1}\log\left(\frac{D(v_i, v_j)+1}{D(v_j)}\right) \quad (5)$$

The dependency on semantic coherence alone, however, might produce meaningless topics dominated by overly ubiquitous words, which are unlikely to capture the unique contents (Bischof and Airoldi, 2012; Airoldi and Bischof, 2016). Therefore, in addition to the semantic coherence, *FREX* (Bischof and Airoldi, 2012) is used as a measure of exclusivity. Mathematically, *f*or word v in topic k, *FREX* is parametrized as:

$$FREX_{k,v} = \left(\frac{w}{ECDF(\beta_{k,v}/\sum_{j=1}^{K}\beta_{j,v})} + \frac{1-w}{ECDF(\beta_{k,v})}\right)^{-1} \quad (6)$$

where $ECDF$ refers to the empirical cumulative distribution function and w is the weight of exclusivity. Following Roberts et al. (2019), w is set at 0.7 in favor of exclusivity. A higher score of *FREX* represents better exclusivity.

All in all, regardless of what diagnostic tools are used, manual review of each topic is indispensable in discerning its semantic interpretations.

2.3 Topic–word assignment

Once the optimal number of topics is determined, the following procedure aims to explore feature words representing a topic. One simplest means is to extract words with the highest probability of presence. This practice, however, tends to produce results biased toward commonly used words that spread across multiple topics. Several refined metrics, such as *FREX*, *Lift*, and *Score*, have therefore been proposed. Unlike *FREX* which calculates the harmonic mean of a word by accounting for both the exclusivity and overall frequency presented in Eq. (6), *Lift* weights a word by dividing by its frequency in other topics, thereby reducing the weight of words that frequently appear in other topics (Taddy, 2013):

$$Lift_{k,v} = \frac{f_{k,v}}{\sum_{k=1}^{K} f_{k,v}/K} \quad (7)$$

where $f_{k,v}$ is the frequency of word v in topic k.

Slightly different from *Lift*, *Score* divides the frequency of word v in topic k by its frequency in other topics after a natural logarithmic transformation:

$$Score_{k,v} = \frac{\log(f_{k,v})}{\sum_{k=1}^{K}\log(f_{k,v})/K} \quad (8)$$

2.4 Topic interpretation and validation

After assigning words to each topic, researchers need to interpret their semantic implications. This is one of the most important steps, as it directly affects subsequent inferences. The aforementioned metrics (i.e., the highest probability, *FREX*, *Lift*, and *Score*) provide a preliminary impression of feature words assigned for each topic. Afterward, by revisiting original document narratives that are estimated to be highly associated with a given topic, researchers can gain a more thorough understanding of textual contexts for the generated topics, by which the quality of topic labeling process can be guaranteed.

3. Data Preparation

Historical data on non-collision injury incidents on public buses were extracted from the Traffic Road Accident Database System maintained by the Hong Kong Police Force and Hong Kong Transport Department (Xu et al., 2019, 2021, 2022; Chen et al., 2022; Zeng et al., 2023; Ye et al., 2024). These non-collision incidents were collected by well-trained police officers alerted by bus operators, bus passengers, or witnesses at the scene. By retrieving information on the type of casualty, vehicle, and collision simultaneously (i.e., casualty type = passenger, vehicle type = public bus, and collision type = non-collision), 13,402 records describing public bus passenger injuries as a result of non-collision incidents during 2010–2019 were extracted. After excluding observations with incomplete information, 12,823 (95.68%) valid samples were retained for analysis. Among these, 89.32% were slight injuries, whereas severe injuries and fatalities accounted for 10.59% and 0.09%, respectively. Here, victims who died immediately or within 30 days of the incident were recorded as fatalities, while those admitted to hospitals for more than or less than 24 hours were counted as the severe or slight injuries, respectively (Zhou et al., 2020).

After trimming extraneous whitespace and special characters, the raw narratives were filtered by two wordlists, i.e., a general stop word list including prepositions, pronouns, articles, and common verbs and a local street name list crawled from OpenStreetMap. As Fig. 1 illustrates, after removing words that were semantically irrelevant, the effective length of narratives under investigation ranged from 8 to 69 words, with the average length being 19 words.

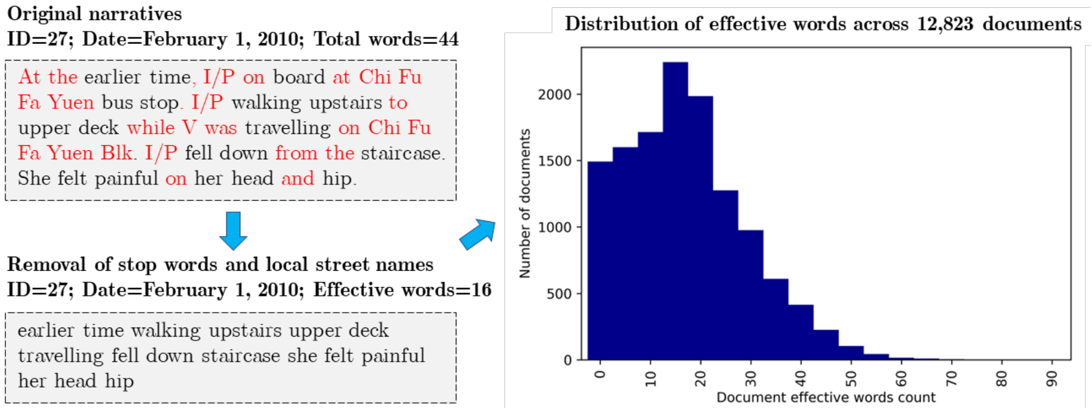

Fig. 1. Illustration of word filtering.

Overall, our corpus contained 6425 unique words. Fig. 2 shows that the top 10 most frequently used words were *lost*, *balance*, *along*, *travel*, *stop*, *fell*, *bus*, *deck*, *lane*, and *reach*, with a frequency of 6974, 6228, 5271, 5134, 5057, 4935, 4926, 4228, 4083, and 3653, respectively.

We then used incident ID to link the extracted narratives with other structured metadata. The metadata used in the present study were the year of incident occurrence and the injury severity of public bus passengers. The estimation of STM, selection of the optimal number of topics, assignment of representative words to each topic, and semantic interpretation of topics were implemented using the freeware R studio (R Core Team, 2019) with the recently released *stm* package (Roberts et al., 2019).

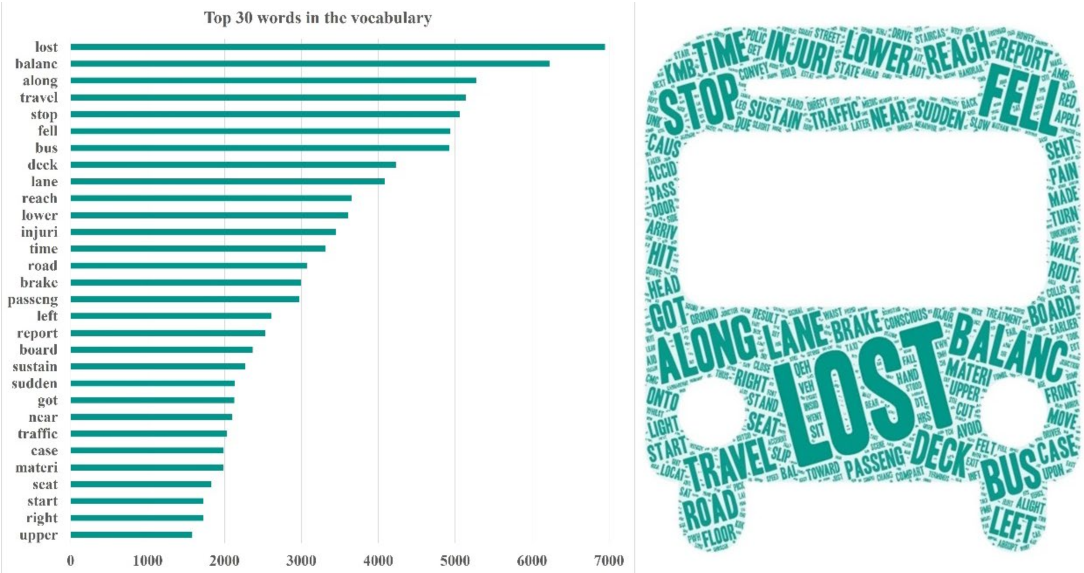

Fig. 2. High-frequency words in the corpus.

4. Results and Discussions

This section sequentially presents and discusses the results of various analyses performed, including the topic selection, topic prevalence, and topic co-occurrence. The discussion highlights particularly how the large-scale but unstructured narratives recorded by the police can be mined by leveraging emerging techniques in natural language processing to elicit the unique characteristics of non-collision injury incidents on public buses.

4.1 Topic selection

The first step was to determine the optimal number of topics given document corpus. Fig. 3 presents the diagnostic results for models with topic numbers ranging from 10 to 100. The left panel shows the relationship between residuals and number of topics, whereas the right part illustrates the plot of semantic coherence against exclusivity. According to Robert et al. (2019), the preferable model is the one with a low residuals value and high scores for semantic coherence and exclusivity. Therefore, the model with 40 topics outperformed, as it yielded the lowest residual value, with relatively higher levels of overall exclusivity and semantic coherence.

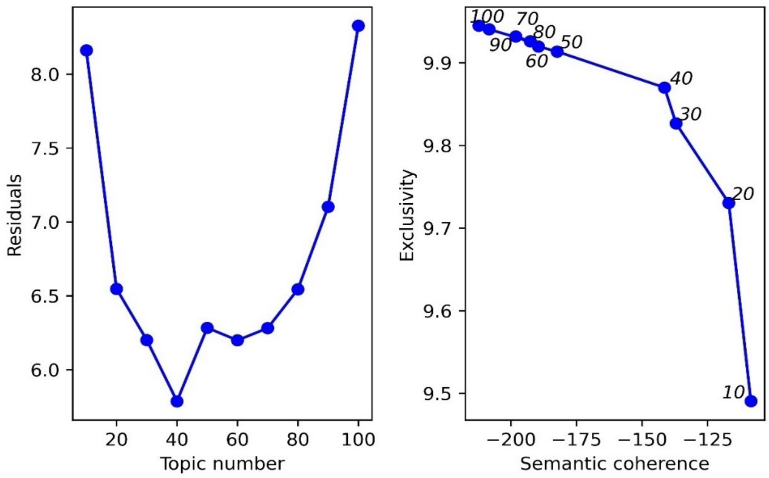

Fig. 3. Results of diagnostic indicators for models with different topics (semantic coherence and exclusivity were computed by Eqs. (5) and (6), respectively).

4.2 Topic labeling and interpretation

The next task was to interpret the thematic structure given a specific number of topics. Data-driven diagnostic measures including the highest probability, *FREX*, *Score*, and *Lift* were first used to identify the most representative words. The context of each topic was then inferred by reviewing the original documents that were most relevant to the given topic. We thereafter presented the generated topics and associated narratives to subject matter experts, including public health specialists, traffic engineers, and public bus operators, to ensure consistent and unambiguous interpretations of topic implications. Evidently, each topic, as presented in Table 1, was deliberately labeled through an integration of representative words, raw narratives, and expert judgements.

Among the 40 generated topics, seven were not successfully labeled. The main reason was that we failed to explicitly infer the content of topics by reviewing the representative words identified by diagnostic measures, given that the feature words of these topics were either semantically meaningless, redundant, or incoherent (e.g., the topic characterized by *drive*, *drove*, and *public*). We therefore removed these seven topics and retained the remainder for analysis.

Table 1. Feature words assigned for each topic.

Topic label	Category[†]	Highest probability	*FREX*	*Lift*	*Score*
NWFB driver	2	driver, found, NWFB	victim, sound, previous	driver, NWFB, eastward	driver, victim, found
door	1	door, passenger, alight	door, open, trap	button, door, open	door, alight, close
passenger alighting	1	passenger, balance, alighting	passenger, balance, avoid	balance, alighting, passenger	balance, passenger, stair
unknown	NA	unknown, change, avoid	unknown, pedestrian, avoid	avoid, unknown, lane	unknown, change, lane
unknown	NA	information, told, hours	information, hours, now	address, alight, told	information, hours, told
unknown	NA	street, short, junction	KMB, years, estate,	chap, NWFB, street	street, estate, years
MTR station	3	station, onboard, MTR	Kowloon, trip, station	fill, Kowloon, trend	station, bay, MTR
traveling along	2	road, travel, along	way, two, toward	convey, along, road	road, toward, travel
unknown	NA	terminus, accord, park	know, office, terminus	argument, awake, know	know, terminus, office
injured person	1	injury, just, person	investigate, proceed, lamp	engage, red-sign, lamp	investigate, injury, person
dashcam	3	took, case, report	took, dashcam, notify	dashcam, barrier, took	took, install, dashcam
stand and lost balance	1	lost, balance, fell	balance, lost, ground	passenger, driving, duck	lost, balance, stand
left/right turn	2	left, right, turn	right, turn, knee	chafe, stretch, surgery	right, turn, left
KMB bus	2	road, KMB, route	weather, KMB, route	head, arm, KMB	route, road, KMB
case handled by	3	case, handle, enquiry	action, enquiry, take	action, attention, log	case, enquiry, action
wheelchair	1	witness, exit, seat	belt, wheelchair, chair	wheelchair, belt, disability	witness, chair, wheel
red traffic light	3	traffic, light, red	light, red, signal	red, signal, traffic	traffic, light, red
approaching station	2	approaching, convey, station	pend, luggage, approaching	airport, baggage, mention	pend, approaching, station
change lanes carelessly	2	lane, travel, left	careless, lane, cut	careless, cut, lane	lane, cut, careless
stopped bus	3	stopped, bus, start	bus, cause, stopped	KMB, stopped, CityBus	stopped, cause, bus
lower-deck seats	1	lower, seat, deck	row, rear, window	handcart, comfort, row	seat, lower, deck
unknown	NA	drive, drove, public,	drive, drove, public	hole, continue, drive	drive, drove, public
ahead condition	2	ahead, condition, time	relevant, injury, condition	offence, walkway, convict	ahead, condition, injury
CCTV	3	alleged, CCTV, intend	prior, list. camera	access, CCTV, footage	access, CCTV, prior
mother	1	east, mother, twist	east, yet, mother	yet, southern, mother	yet, east, mother
ambulance arrived	3	arrive, made, ambulance	report, hospital, ambulance	hospital, uptown, lost	ambulance, arrive, made
minor injury	1	injury, passenger, conscious	injury, sustain, minor	conscious, passenger, move	passenger, injury, minor
unknown	NA	claim, body, run	run, carry, turn	run, construct, site	claim, run, body
nearside pedestrian	3	cross, nearside, pedestrian	pedestrian, cross, nearside	auto, bell, alarm	cross, nearside, pedestrian
bus terminus	3	bus, board, stop	bus, terminus, board	alight, hospital, chest	bus, board, terminus
driving inattentively	2	lose, drive, inattentive	inattentive, code, road	inattentive, road, dissatisfied	inattentive, lose, drive
feel painful and report later	1	report, felt, pain	sought, felt, later	felt, treatment, report	report, felt, later
vehicle slips	2	travel, slip, along	vehicle, destination, slip	slip, vehicle, destination	vehicle, slip, earlier
sudden/hard brake	2	brake, sudden, front	hard, brake, prevent	collision, brake, wind	brake, hard, sudden
fail to hold handrail	1	handrail, hold, hand	hold, tight, firm	Bag, firm, grab	handrail, hold, fail
roundabout	3	taxi, roundabout, front	outer, taxi, circle	circle, cyclist, outer	taxi, white, roundabout
walk upstairs	1	deck, upper, walk	upper, staircase, upstairs	grasp, junction, balance	upper, staircase, walk
fall onto floor	1	time, fall, floor	location, floor, fall	collision, somehow, fall	fall, floor, location
avoid collision	2	avoid, collision, abrupt	collision, abrupt, avoid	disappear, incorrect, unexpected	avoid, collision, abrupt
unknown	NA	reach, near, compartment	compartment, reach, near	wait, compartment, reach	wait, reach, compartment

Note: NWFB, KMB, MTR, and CCTV refer to New World First Bus Company, Kowloon Motor Bus Company, MTR Corporation Limited, and closed-circuit television, respectively.
Categories labeled as 1, 2, and 3 represent passenger-related, bus-related, and incident-related characteristics, respectively.

Following Zhou et al. (2020), we further stratified the topics of interest as three broad categories, which described the characteristics of injured passengers, buses, and incidents, respectively. Table 1 shows that 12 topics were determined as passenger-related characteristics, because these topics were closely related to locations (i.e., topics labeled *door* and *lower-deck seats*), actions (i.e., topics labeled *passenger alighting*, *stand and lost balance*, *fail to hold handrail*, *walk upstairs*, and *fall onto floors*), injury outcomes (i.e., topics labeled *feel painful and report later, injured person*, and *minor injury*), and identities (i.e., topics labeled *wheelchair* and *mother*) of injured passengers. Likewise, topics indicative of bus driver identities (i.e., *NWFB driver* and *KMB bus*), bus driver maneuvers (i.e., *change lanes carelessly*, *ahead condition*, *driving inattentively*, *sudden/hard brake*, and *avoid collision*), and bus operating states (i.e., *traveling along*, *left/right turn*, *approaching station*, *stopped bus*, and *vehicle slips*) were categorized as bus-related characteristics.

In addition, nine out of 33 topics were judged as incident-related characteristics, because these topics briefly described the locations of incidents (i.e., topics entitled *MTR station*, *red traffic light*, *bus terminus*, and *roundabout*), information sources (i.e., topics labeled *dashcam* and *CCTV*), involved third-parties (i.e., topic namely *nearside pedestrians*), and first-aid responses (i.e., topics labeled *case handled by* and *ambulance arrived*). Compared with previous studies (Silvano and Ohlin, 2019; Siman-Tov et al., 2019; Zhou et al., 2020) which depended on predefined variables collected in tabulated forms, our generated topics portray a more holistic picture of non-collision injury incidents on public buses, which has not previously been reported.

4.3 Topic prevalence

Fig. 4 illustrates the overall prevalence of topics across all the samples over the studied period. Interestingly, the topics with the highest proportions were mainly those that described the pre-injury actions of passengers (i.e., *stand and lost balance*), bus company (i.e., *KMB bus*), operating states of buses (i.e., *vehicle slips*), and inappropriate maneuvers of bus drivers (i.e., *change lanes carelessly*), followed by the topics labeled *walk upstairs*, *fall onto floors*, *feel painful and report later*, *door*, and *lower-deck seats*. All these topics with higher rankings pertain to either passenger- or bus-related characteristics. Likewise, the three most prevalent topics in relation to incident characteristics were *ambulance arrived*, *bus terminus*, and *red traffic light*. It is worth mentioning that, as the most dominant topic with a proportion as high as 14.8%, the *stand and lost balance* topic generally expresses two main implications. First, it indicates the condition of passengers at the time of injury, which is standing on the bus instead of sitting on a seat, climbing stairs, boarding, or alighting from the bus. Second, it describes the cause of bus passengers being injured, highlighting particularly the importance of keeping balance, e.g., by holding handrails while standing on the bus. This finding is consistent with Silvano and Ohlin (2019) and Siman-Tov et al. (2019), who found that standing passengers were overrepresented among non-collision injuries on public buses, primarily because they were more frequently subject to acceleration and braking maneuvers. An elaborate biomechanical experiment conducted by Karekla and Tyler (2018) also demonstrated that standing passengers struggled to maintain balance if a bus accelerated at a rate of 2.0 m/s^2 or above.

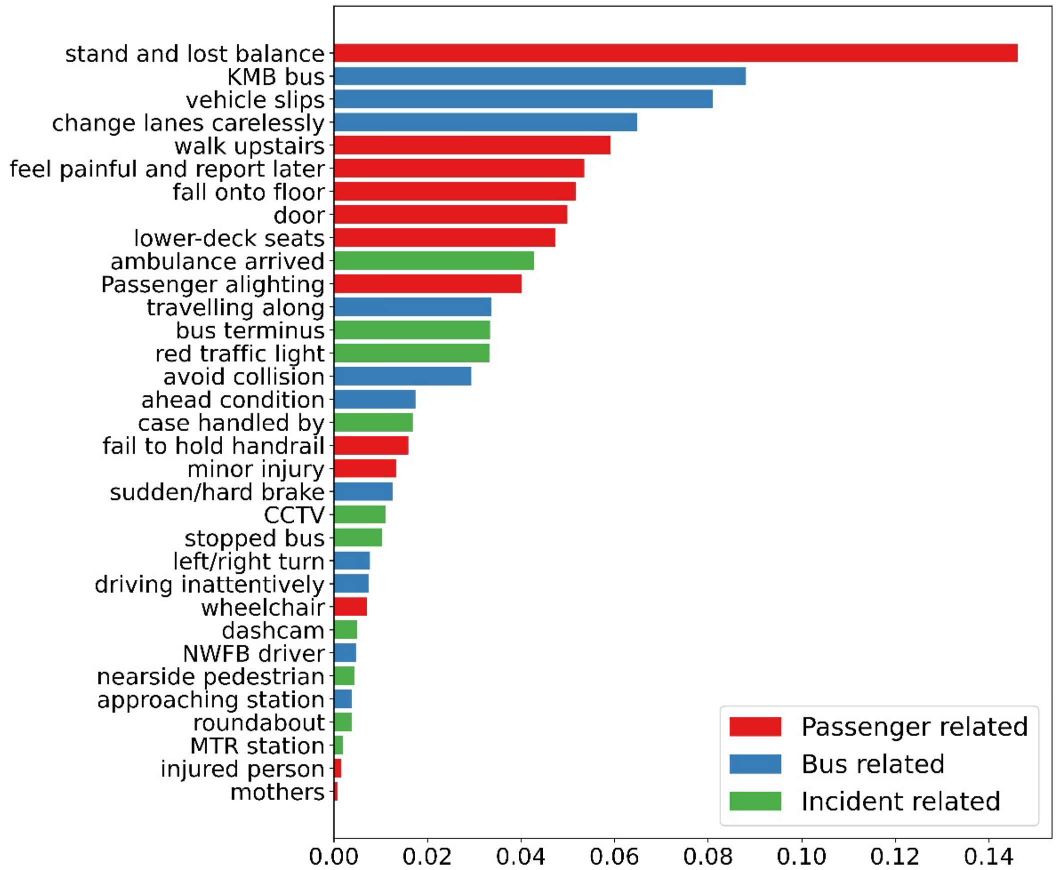

Fig. 4. Graphical display of the topic prevalence (NWFB, KMB, MTR, and CCTV refer to the New World First Bus Company, Kowloon Motor Bus Company, MTR Corporation Limited, and closed-circuit television, respectively).

4.3.1 Topic prevalence stratified by injury outcomes

One major advantage of STM is the ability to easily link topics with metadata. Given the inherent differences in risk factors associated with injury severity among public bus passengers (Zhou et al., 2020), the prevalence of topics is expected to vary substantially across incidents with different injury outcomes. We thereby stratified the topic proportion by slight or fatal/severe injuries. The combination of fatalities with severe injuries here was unlikely to affect inferences, as fatal incidents accounted for far less than 1% of our sample.

Fig. 5 presents topic prevalence stratified by the severity of injuries sustained by public bus passengers in non-collision incidents. Broadly, the ranking of topics in narratives resulting in fatal and severe injuries was similar to that of slight injuries. A closer look at topic proportions across these two categories, however, indicates subtle discrepancies. Specifically, topics namely *vehicle slips*, *walk upstairs*, and *ambulance arrived* were more dominant in incidents with fatal and severe injuries, as they ranked relatively higher (i.e., the second, third, and fifth, respectively) in the right panel of Fig. 5.

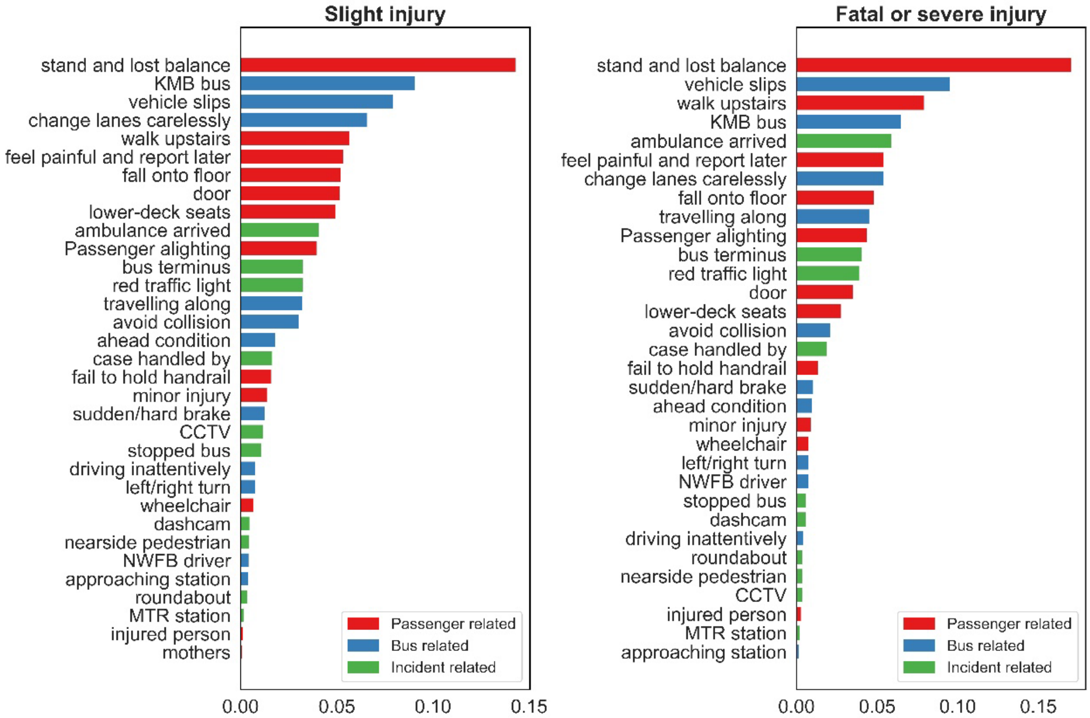

Fig. 5. Topic prevalence stratified by the injury severity of onboard passengers (NWFB, KMB, MTR, and CCTV refer to the New World First Bus Company, Kowloon Motor Bus Company, MTR Corporation Limited, and closed-circuit television, respectively).

Based on the feature words associated with each topic, 27 unique variables were constructed through content analysis and keyword extraction (Wali et al., 2021). Specifically, topics such as *injured person*, *feel painful and report later*, *minor injury*, *case handle by*, and *ahead condition* failed to generate new variables, either because of semantic redundancy or lack of semantic integrity. A Chi-square test was conducted to investigate whether the distribution of contextual variables differed significantly from injury outcomes. To quantify the effects of various factors, odds ratios (Zeng et al., 2023) were estimated by the fixed-parameter logistic regression model, as none of the explanatory variables resulted in significantly heterogeneous effects in the random-parameter model. The results are presented in Table 2.

As table 2 indicates, seven factors were significantly associated with the severity of injuries to public bus passengers in non-collision incidents. Consistent with the findings of Siman-Tov et al. (2019), standing and boarding passengers sustained a substantially higher likelihood of fatal and severe injuries, with the odds increasing by 34% and 37%, respectively. Similarly, when a bus skidded, the odds of passengers being fatally and severely injured increased by 38%. Zhou et al. (2020) also reported that bus passengers were more likely to suffer from fatal and severe injuries when non-collision incidents occurred during heavy rain. This result is expected because wet road surfaces greatly reduce friction, which may lead to a loss of vehicle control such as skidding and brake failure.

Interestingly, the odds of fatal and severe injuries decreased by as much as 72% when a mother with a child was injured on a public bus. This reduced likelihood may be attributable to risk compensation. That is, the mother tends to be more cautious with the child when taking a bus, which helps decrease the risk of serious injury outcomes.

Passengers not holding handrails when boarding, standing on the deck, or alighting experienced a 34% decrease in the odds of fatal and severe injuries. This counterintuitive result arises likely from the absence of standardized and uniform guidelines for police officers to write narratives (Lopez et al., 2022), along with potential underreporting (Abay, 2015) or missing imperative incident details (Ahmed et al., 2019). We therefore call for future studies to leverage data derived from other sources, such as the vehicle-mounted videos, to validate this finding.

Table 2. Effects of contextual variables derived from STM on the injury severity of public bus passengers in non-collision incidents.

Variables	Coding	Fatal or severe injury	Slight injury	*p*-value	Unadjusted odds ratio		Adjusted odds ratio	
					Mean	95% CI	Mean	95% CI
Passenger related								
Standing and losing balance	Yes	747 (11.82%)	5574 (88.18%)	0.00**	1.30**	(1.16, 1.45)	**1.34****	**(1.19, 1.50)**
	No†	609 (9.37%)	5893 (90.63%)					
Walking upstairs	Yes	32 (13.97%)	197 (86.03%)	0.09*	1.38*	(0.95, 2.02)	1.26	(0.86, 1.84)
	No†	1324 (10.51%)	11,269 (89.49%)					
Falling onto floor	Yes	85 (12.21%)	611 (87.79%)	0.15	1.19	(0.94, 1.50)	1.15	(0.91, 1.46)
	No†	1271 (10.48%)	10,856 (89.52%)					
Boarding	Yes	75 (14.15%)	455 (85.85%)	0.01**	1.42**	(1.10, 1.82)	**1.37****	**(1.06, 1.76)**
	No†	1281 (10.42%)	11,012 (89.58%)					
Alighting from the bus	Yes	125 (11.15%)	996 (89.43%)	0.51	1.07	(0.88, 1.30)	1.10	(0.90, 1.34)
	No†	1231 (10.52%)	10,471 (89.48%)					
Seated on lower deck	Yes	82 (9.50%)	781 (90.50%)	0.29	0.88	(0.70, 1.11)	0.91	(0.72, 1.16)
	No†	1274 (10.65%)	10,686 (89.35%)					
Failure to hold handrail	Yes	57 (7.14%)	741 (92.86%)	0.00**	0.64**	(0.48, 0.84)	**0.66****	**(0.50, 0.87)**
	No†	1299 (10.80%)	10,726 (89.20%)					
Wheelchair	Yes	5 (13.89%)	31 (86.11%)	0.52	1.37	(0.53, 3.52)	1.47	(0.57, 3.82)
	No†	1351 (10.57%)	11,436 (89.43%)					
Mothers with children	Yes	4 (3.08%)	126 (96.92%)	0.01**	0.27**	(0.10, 0.72)	**0.28****	**(0.10, 0.76)**
	No†	1352 (10.65%)	11,341 (89.35%)					
Bus related								
KMB bus	Yes	158 (8.05%)	1804 (91.95%)	0.00**	0.71**	(0.59, 0.84)	**0.73****	**(0.61, 0.87)**
	No†	1198 (11.03%)	9663 (88.97%)					
NWFB bus	Yes	32 (11.94%)	236 (88.06%)	0.46	1.15	(0.79, 1.67)	1.11	(0.76, 1.62)
	No†	1324 (10.55%)	11,231 (89.45%)					
Vehicle slips	Yes	136 (13.57%)	866 (86.43%)	0.00**	1.36**	(1.13, 1.65)	**1.38****	**(1.14, 1.68)**
	No†	1220 (10.32%)	10,601 (89.68%)					
Bus stopped	Yes	14 (9.72%)	130 (90.28%)	0.74	0.91	(0.52, 1.58)	0.83	(0.47, 1.45)
	No†	1342 (10.58%)	11,337 (89.42%)					

Traveling along	Yes	141 (9.78%)	1301 (90.22%)	0.30	0.91	(0.75, 1.09)	0.93	(0.77, 1.12)
	No†	1215 (10.68%)	10,166 (89.32%)					
Turning left/right	Yes	11 (7.64%)	133 (92.36%)	0.25	0.70	(0.38, 1.29)	0.69	(0.37, 1.28)
	No†	1345 (10.61%)	11,334 (89.39%)					
Approaching station	Yes	12 (10.34%)	104 (89.66%)	0.94	0.98	(0.54, 1.78)	0.96	(0.52, 1.75)
	No†	1344 (10.58%)	11,363 (89.42%)					
Sudden/hard brake	Yes	37 (10.95%)	301 (89.05%)	0.82	1.04	(0.74, 1.47)	1.11	(0.78, 1.58)
	No†	1319 (10.56%)	11,166 (89.44%)					
Changing lane carelessly	Yes	25 (7.18%)	323 (92.82%)	0.04**	0.65**	(0.43, 0.98)	0.76	(0.50, 1.15)
	No†	1331 (10.67%)	11,144 (89.33%)					
Driving inattentively	Yes	3 (5.00%)	57 (95.00%)	0.16	0.44	(0.14, 1.42)	0.49	(0.15, 1.60)
	No†	1353 (10.60%)	11,410 (89.40%)					
Avoiding collisions	Yes	45 (6.44%)	654 (93.56%)	0.00**	0.57**	(0.42, 0.77)	**0.62****	**(0.46, 0.85)**
	No†	1311 (10.81%)	10,813 (89.19%)					
Incident related								
Ambulance arrived	Yes	4 (7.41%)	50 (92.59%)	0.45	0.68	(0.24, 1.87)	0.67	(0.24, 1.88)
	No†	1352 (10.59%)	11,417 (89.41%)					
Bus terminus	Yes	53 (11.21%)	420 (88.79%)	0.65	1.07	(0.80, 1.43)	1.05	(0.78, 1.41)
	No†	1303 (10.55%)	11,047 (89.45%)					
MTR station	Yes	8 (11.59%)	61 (88.41%)	0.78	1.11	(0.53, 2.32)	1.25	(0.59, 2.63)
	No†	1348 (10.57%)	11,406 (89.43%)					
Roundabout	Yes	12 (10.81%)	99 (89.19%)	0.94	1.03	(0.56, 1.87)	1.03	(0.56, 1.90)
	No†	1344 (10.57%)	11,368 (89.43%)					
Red traffic light	Yes	37 (11.04%)	298 (88.96%)	0.78	1.05	(0.74, 1.49)	1.08	(0.76, 1.54)
	No†	1319 (10.56%)	11,169 (89.44%)					
CCTV	Yes	8 (6.40%)	117 (89.43%)	0.13	0.58	(0.28, 1.18)	0.59	(0.29, 1.22)
	No†	1348 (10.62%)	11,350 (89.38%)					
Dashcam	Yes	10 (12.66%)	69 (87.34%)	0.55	1.23	(0.63, 2.39)	1.22	(0.63, 2.39)
	No†	1346 (10.56%)	11,398 (89.44%)					

†: reference category; **: significant at 95% confidence level; *: significant at 90% confidence level. All extracted variables were included during the estimation of adjusted odd ratios, because of the absence of strong collinearity.

Lastly, our study found that unlike improper driving drivers such as careless lane changing and inattentive driving, passengers were less likely to be fatally and severely injured in non-collision incidents if the bus driver was attempting to avoid a collision with other road users such as nearside pedestrians. One plausible explanation is that public bus drivers in Hong Kong, particularly those employed by the KMB, are well-trained to prioritize passenger safety in emergency situations such as collision avoidance (Chen et al., 2022; Loo et al., 2023). It is thereby unsurprising that the odds of fatal and severe injuries decreased by 27% for passengers on KMB buses.

4.3.2 Topic prevalence stratified by year of occurrence

To further uncover the temporal variations in topic prevalence across the period of interest, by stratifying the topics by temporal variable, the yearly dynamic of topic prevalence during 2010–2019 was profiled. As Fig. 6 shows, the left panel presents the cross-sectional distribution of topics over the observation period (i.e., the sum of topic proportions in a specific year is equal to 1), whereas the right panel illustrates the longitudinal imbalance of each topic (i.e., the sum of a specific topic proportion across the studied period is equal to 1). By assuming that each topic accounts for 10% each year if equally distributed over the 10-year period, the longitudinal imbalance of topics can then be quantified by the difference between the observed and expected values.

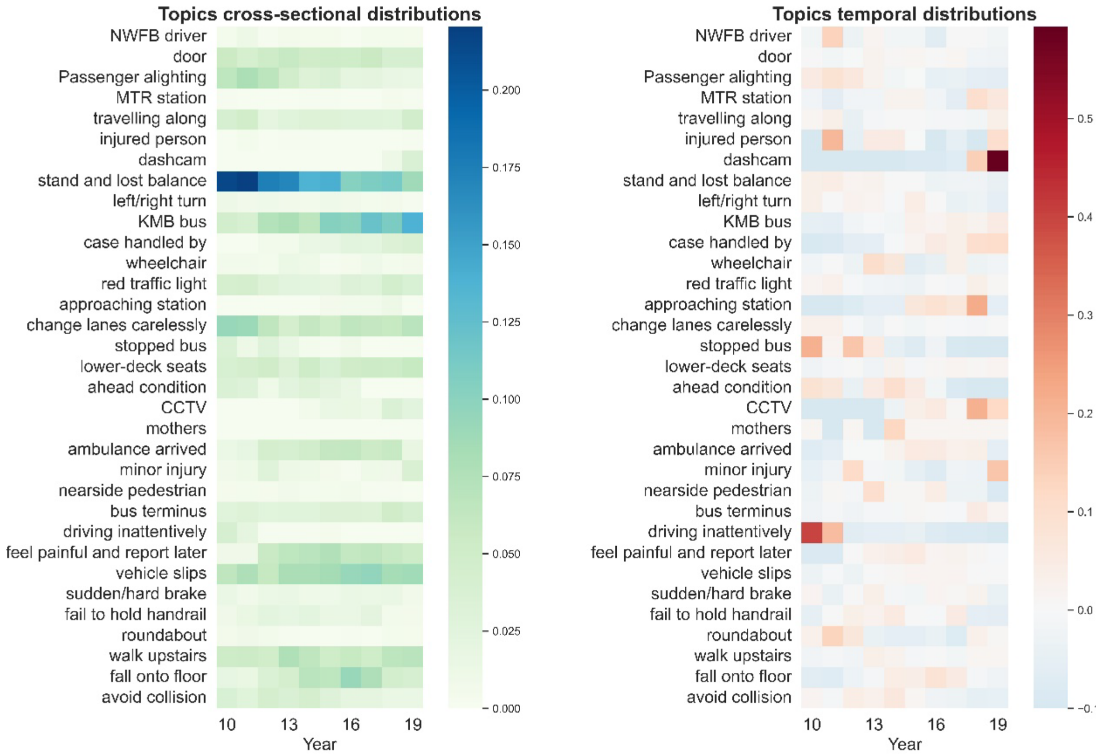

Fig. 6. Dynamics of topic prevalence (the left side shows the cross-sectional distribution over the 10-year period, whereas the right side presents the longitudinal imbalance. NWFB, KMB, MTR, and CCTV refer to the New World First Bus Company, Kowloon Motor Bus Company, MTR Corporation Limited, and closed-circuit television, respectively).

According to the left panel of Fig. 6, the topic labeled *stand and lost balance* played the most dominant role over all topics but became increasingly less prevalent. One plausible explanation is that public transit service in Hong Kong has improved (Tong and Ng, 2021) that fewer passengers have to stand during their journeys. For another, passengers might raise their safety awareness, e.g., by grasping handrails while walking or standing on the bus. Conversely, the share of topic labeled *KMB bus* increased steadily from 2% in 2010 to 13% in 2019. Actions such as initiation of education programs to bus drivers and passengers are needed to improve the safety performance of public buses operated by the Kowloon Motor Bus Company.

Thanks to the untiring promotion of intelligent public transit systems in Hong Kong since 2016 (Chen et al., 2022), the dashcam is now widely used as a reliable source of information by police to investigate non-collision injury incidents on public buses. This fact is well reflected by our findings that the imbalance of the topic labeled *dashcam* increased dramatically, particularly during the last two years, as illustrated in the right panel of Fig. 6. In contrast, topic labeled *driving inattentively* has become gradually inappreciable since 2012, resulting in a more balanced pattern. This result suggests that the occurrence of non-collision injury incidents on public buses due to inattentive driving has reduced substantially.

4.4 Topic co-occurrence

Given the 33 topics generated by STM, a network topology comprising nodes and links was constructed to visualize the intricate relationship between inferred topics. Here, the nodes denote inferred topics, whereas the links between nodes describe the strength of associations, which were computed based on the co-occurrence of words between two topics using the Pearson correlation matrix (Kwayu et al., 2021). The results are illustrated in Fig. 7.

As Fig. 7 shows, the top five links connecting two topics were: 1) *passenger alighting* and *stopped bus*, 2) *sudden/hard brake* and *change lanes carelessly*, 3) *stand and lost balance* and *stopped bus*, 4) *avoid collision* and *change lanes carelessly*, and 5) *avoid collision* and *sudden/hard brake*. Such associations between topics might help clarify the causes of non-collision injury incidents on public buses. For example, the direct and strong connection between topics labeled *avoid collision* and *sudden/hard brake* indicates that non-collision incidents occurred because of the emergency braking maneuvers of bus drivers in an attempt to avoid collisions. Likewise, the topic labeled *passenger alighting* co-occurred repeatedly with *stopped bus*, implying that passengers were injured when alighting from the bus, being hit by the doors or falling because of slippery floors. Another interesting finding is that bus-related topics such as the *KMB bus*, *vehicle slips*, *left/right turn*, and *driving inattentively* were located more centrally, as these topics were more accessibly connected with others.

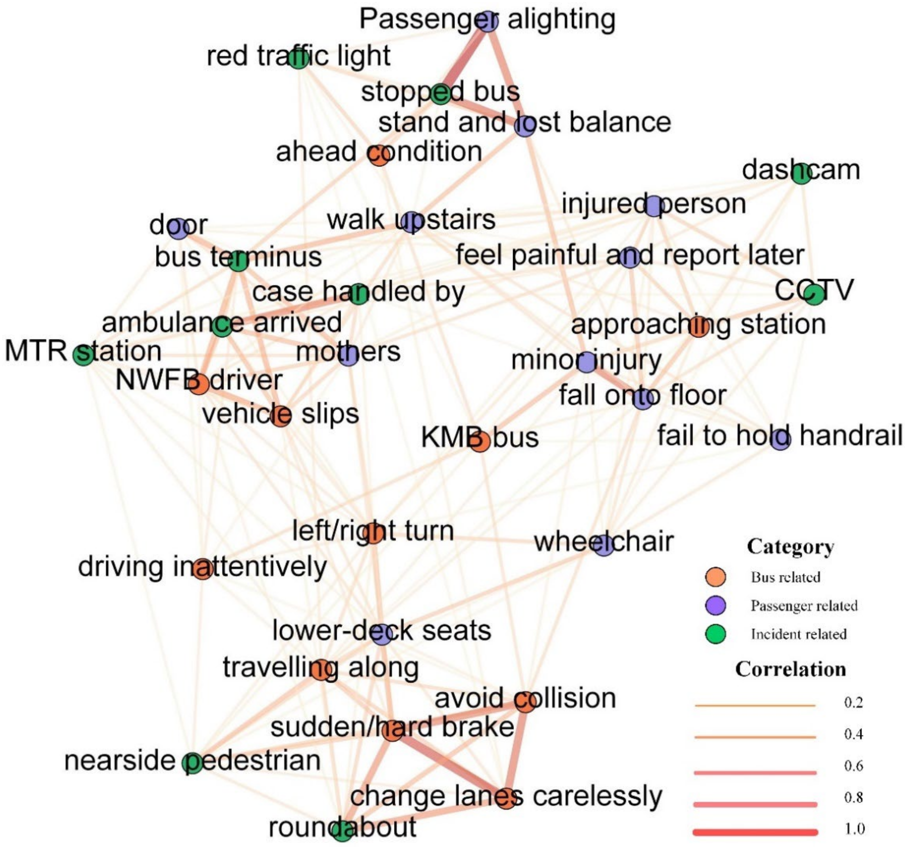

Fig. 7. Network topology of topic co-occurrence (NWFB, KMB, MTR, and CCTV denote the New World First Bus Company, Kowloon Motor Bus Company, MTR Corporation Limited, and closed-circuit television, respectively).

In addition to the pairwise association, we applied the modularity analysis (Chang et al., 2019) to figure out topics that were more likely to simultaneously appear in the same narratives. The results are presented in Table 3 and Fig. 8.

Four clusters emerged with strong inner connections. **Cluster 1** was characterized by topics labeled *fall onto floor*, *fail to hold handrail*, *minor injury*, *feel painful and report later*, *injured person*, *KMB bus*, *driving inattentively*, *approaching station*, *CCTV*, and *dashcam*, with the topic entitled *fall onto floor* being the most critical node of this modularity given the highest weighted degree (i.e., a value of 1.655). This cluster seems to describe non-collision incidents involving passenger falls on public buses operated by the Kowloon Motor Bus Company, potentially resulting from the failure of passengers to hold handrails and the inattentive behaviors of bus drivers when approaching bus stations. Such type of incidents was more likely to result in slight injuries, given the prevalence of topics labeled *minor injury* and *feel painful and report later*.

Table 3. Results of modularity analysis.

Topic label	Role	Cluster	Weighted degree
fall onto floor	Passenger related	1	1.655
fail to hold handrail	Passenger related	1	0.396
minor injury	Passenger related	1	1.570
feel painful and report later	Passenger related	1	1.083
injured person	Passenger related	1	1.080
KMB bus	Bus related	1	0.770

driving inattentively	Bus related	1	0.432
approaching station	Bus related	1	1.206
CCTV	Incident related	1	0.830
dashcam	Incident related	1	0.432
stand and lost balance	Passenger related	2	1.803
passenger alighting	Passenger related	2	1.409
door	Passenger related	2	0.682
ahead condition	Bus related	2	0.553
stopped bus	Bus related	2	1.694
red traffic light	Incident related	2	0.548
mother	Passenger related	3	1.706
walk upstairs	Passenger related	3	1.322
vehicle slips	Bus related	3	1.634
NWFB driver	Bus related	3	1.569
bus terminus	Incident related	3	2.155
MTR station	Incident related	3	0.713
ambulance arrived	Incident related	3	1.611
case handled by	Incident related	3	1.101
lower-deck seats	Passenger related	4	1.696
wheelchair	Passenger related	4	0.805
sudden/hard brake	Bus related	4	2.393
change lanes carelessly	Bus related	4	1.991
avoid collision	Bus related	4	1.963
traveling along	Bus related	4	1.513
left/right turn	Bus related	4	1.269
roundabout	Incident related	4	1.650
nearside pedestrian	Incident related	4	1.119

Topics labeled *stand and lost balance*, *ahead condition*, *red traffic light*, *passenger alighting*, *door*, and *stopped buses* formulated **Cluster 2**, suggesting two scenarios typically associated with non-collision incidents on public buses. One might involve the loss of balance among standing passengers, owing to the sudden and sharp decelerations of bus drivers in response to red traffic lights ahead. This type of non-collision incidents is quite common, particularly among public buses operating in highly urbanized areas with dense road networks like Hong Kong (Zhou et al., 2020). The other pertains to passengers being hit by doors when alighting from the bus, probably because of the early discharges of bus drivers. Such an incident type is well supported by the strong connection of topics labeled *passenger alighting* and *stopped bus*, as illustrated in Fig. 7. As highlighted by Silvano and Ohlin (2019) and Siman-Tov et al. (2019), great stress resulting from a tight schedule might compel bus drivers to pull away from bus stations without waiting sufficient time for passengers to disembark.

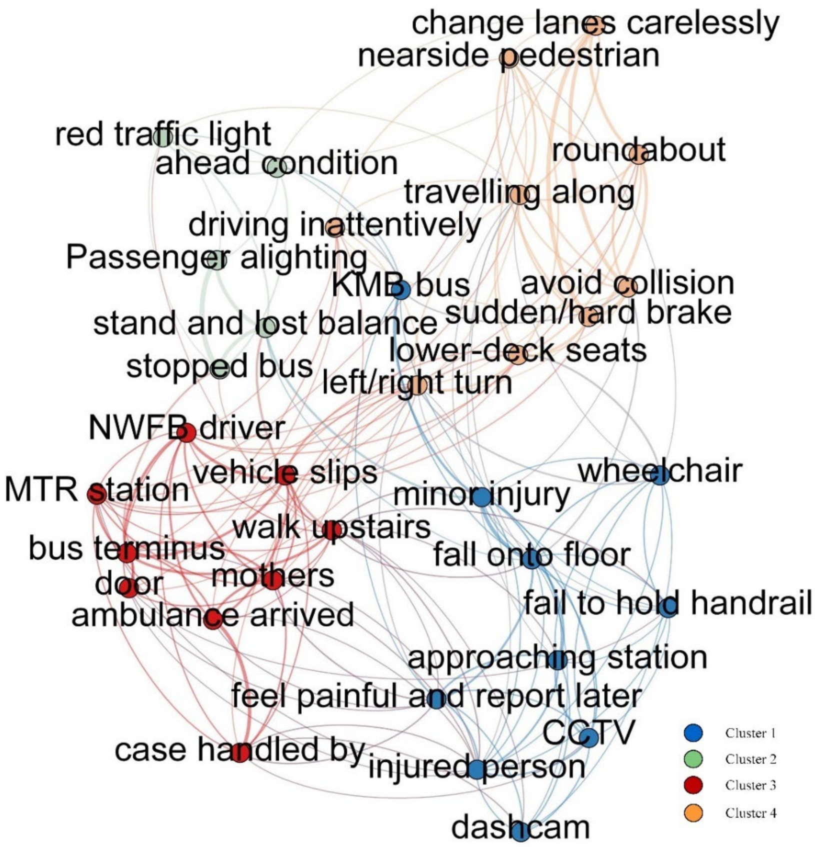

Fig. 8. Modularity analysis of topic co-occurrence (NWFB, KMB, MTR, and CCTV refer to the New World First Bus Company, Kowloon Motor Bus Company, MTR Corporation Limited, and closed-circuit television, respectively).

Interestingly, **Cluster 3** constituted eight topics, namely *mother*, *walk upstairs*, *vehicle slips*, *NWFB driver*, *bus terminus*, *MTR station*, *ambulance arrived*, and *case handled by*. This cluster primarily portrayed non-collision incidents in which passengers were injured when climbing stairs, on buses operated by the New World First Bus Compony, near MTR stations or bus terminus, and because of vehicle skidding. Caution should be paid in particular to this type of incidents, given the prevalence of vulnerable bus commuters (i.e., mothers with children) and more serious injury outcomes (i.e., injured passengers were transferred to hospitals as indicated by the topic labeled *ambulance arrived*).

Finally, **Cluster 4** comprised nine topics labeled *wheelchair*, *lower-deck seats*, *sudden/hard brake*, *change lanes carelessly*, *avoid collision*, *traveling along*, *left/right turn*, *roundabout*, and *nearside pedestrian*. Specifically, disabled passengers with wheelchairs and passengers seated on the lower deck of a bus were overrepresented in this cluster, probably because of the sudden and hard braking of bus drivers in an effort to avoid collisions with nearside pedestrians or due to the careless lane-changing behaviors (e.g., abrupt turning) of bus drivers when weaving through roundabouts. Indeed, emergency braking is most likely to trigger non-collision injury incidents on public buses, as it is usually abrupt and unexpected, leaving little time for passengers to react.

4.5 Implications

Non-collision injuries to public bus passengers are undoubtedly evitable and preventable (Elvik, 2019). Based on the aforementioned findings, tailor-made countermeasures are proposed to reduce non-collision injury incidents on public buses following the "4E" principle (i.e., engineering, education, enforcement, and emergency). Such safety measures might involve setting exclusive bus lanes to reduce conflicts with other road users, redesigning bus stops to reserve more room for public buses to pull in and out (Akintayo and Adibeli, 2022), prohibiting standing too close to doors, restricting the number of standing passengers, improving emergency treatments, and providing systematic training to public bus drivers to enhance their performance in handling urgent situations such as the loss of control of the vehicle, collision avoidance with pedestrians, and driving in heavy rain. To better cater for the needs of vulnerable bus commuters, particularly parents holding babies, the disabled, and the elderly, onboard facilities should be adjusted accordingly, e.g., by promoting the use of soft and textured floors (Halpern et al., 2005), lowering bus steps to improve accessibility (Siman-Tov et al., 2019), and replacing horizontal handholds with vertical ones near doors (Palacio et al., 2009).

In addition, both bus drivers and passengers should raise their safety awareness and foster a greater appreciation of scenarios that are more likely to cause non-collision injuries on public buses. For example, keeping in mind that passengers seated on the upper deck, the elderly, parents with children, and the disabled require more time to alight from a bus, public bus drivers are likely to be more considerate of these passengers and avoid early departures from bus stops. Likewise, being aware that more muscle strength is required to maintain balance when climbing staircases, passengers should be more cautious when walking upstairs by grasping handrails tightly (Karekla and Tyler, 2019). Finally, to ensure the safe operation of public buses, it might be beneficial to deploy advanced driver assistance techniques (Yue et al., 2020), e.g., a holographic sensing system with the ability to proactively identify cutting-in vehicles and nearside pedestrians. An active voice broadcast system can also help remind standing passengers to hold handrails when the bus swerves away from stations, approaches signalized intersections, and weaves through roundabouts.

5. Conclusions

Based on a comprehensive dataset of 12,823 narratives recorded by police during 2010–2019 in Hong Kong, we uncovered the underlying themes of non-collision injury incidents on public buses, revealed their dynamic patterns, and portrayed their co-occurrences by leveraging emerging natural language processing techniques. Unlike Alambeigi et al. (2020), Kwayu et al. (2021), and Rose et al. (2022) who only identified latent topics without further exploring their intricate interactions, by integration of STM and network topology analysis, our study provides a more panoramic view of public bus passenger injuries as a result of non-collision incidents, which helps deduce causation chains for different incident types.

Several key findings are worth mentioning. First, standing passengers were overrepresented in non-collision injury incidents on public buses, given the dominant role of the topic labeled *stand and lost balance*. Second, public bus passengers were more likely to be fatally and severely injured, when the non-collision incidents occurred because the bus skidded, when passengers were boarding, and when standing passengers lost their balance. Third, by fusing STM with network topology analysis, we provide new insights by figuring out topics that co-occur frequently. Specifically, six unique patterns associated with non-collision injury incidents on public buses were untangled, that is, the failure to hold handrails accompanied with the inappropriate behaviors of bus drivers when approaching bus stations, the loss of balance among standing passengers due to the sudden and sharp braking of bus drivers in response to red traffic lights ahead, passengers being hit by the door when alighting from a bus, passengers falling while climbing stairs to the upper deck, passengers being injured because of the emergency actions of bus drivers to avoid collisions with nearside pedestrians, and passengers being injured due to the careless lane changing of bus drivers when weaving through roundabouts. These scenarios have not yet been reported and should serve as a foundation for the formulation of evidence-based safety measures. Like Kwayu et al. (2021), Kutela et al. (2022b), and Jing et al. (2023), although the results mined by STM is primarily explanatory in nature, future studies can benefit from our study when designing biomechanical experiments to investigate the determinants of passenger injuries resulting from non-collision incidents on public buses.

This study is not without limitations. Unlike epidemiological studies based on data retrieved from hospital admission reports (Björnstig et al., 2005; Halpern et al., 2005; Zunjic et al., 2012; Silvano and Ohlin, 2019; Siman-Tov et al., 2019; Chen et al., 2024), our non-collision incidents on public buses were regularly collected and complied by the police. As the sole representative and reliable data source publicly available over such a long timeslot in Hong Kong, these police reports have been routinely used by local authorities for decision making (Xu et al., 2019; 2022; Zhou et al., 2020; Chen et al., 2022; Zeng et al., 2023). The unambiguous and accordant interpretations of topic implications among subject matter experts further demonstrate that large-scale, unstructured textual narratives recorded by the police can serve as a valuable and organized information source for cause analysis by harnessing state-of-art natural language processing techniques. Additional studies with newly released injury reports from other regions are highly advocated to validate our findings. One fundamental assumption associated with STM is the bag-of-words (Grimmer and Stewart, 2013), which simplifies the raw narratives as collections of words without taking word sequence into account. In this regard, *bus stopped* and *stopped bus* share the same representation in bag-of-words models despite their slightly different semantics. Researchers can leverage more advanced text-vectorization methods such as word embeddings to better capture semantic relationships (Goldberg, 2022; Liu and Yang, 2022).

References

Abay, K.A., 2015. Investigating the nature and impact of reporting bias in road crash data. *Transportation Research Part A: Policy and Practice* 71, 31–45.

Adämmer, P., Schüssler, R.A., 2020. Forecasting the equity premium: mind the news! *Review of Finance* 24(6), 1313–1355.

Ahmed, A., Sadullah, A.F.M., Yahya, A.S., 2019. Errors in accident data, its types, causes and methods of rectification-analysis of the literature. *Accident Analysis & Prevention* 130, 3–21.

Airoldi, E.M., Bischof, J.M., 2016. Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association* 111(516), 1381–1403.

Akintayo, F.O., Adibeli, S.A., 2022. Safety performance of selected bus stops in Ibadan Metropolis, Nigeria. *Journal of Public Transportation* 24, 100003.

Alambeigi, H., McDonald, A.D., Tankasala, S.R., 2020. Crash themes in automated vehicles: a topic modeling analysis of the California Department of motor vehicles automated vehicle crash database. arXiv preprint arXiv:2001.11087.

Aminpour, N., Saidi, S., 2025. Unveiling mobility patterns beyond home/work activities: A topic modeling approach using transit smart card and land-use data. *Travel Behaviour and Society* 38, 100905.

Arabian, A., Masjoodi, S., Makkiabadi, B., Ghafari, E., Nassaj, E.T., Zakerian, S.A., 2020. Determination of critical time points in non-collision incidents of elderly passengers in standing position on urban bus. *Traffic Injury Prevention* 21(2), 151–155.

Arteaga, C., Paz, A., Park, J., 2020. Injury severity on traffic crashes: a text mining with an interpretable machine-learning approach. *Safety Science* 132, 104988.

Baburajan, V., de Abreu e Silva, J., Pereira, F.C., 2022. Open vs closed-ended questions in attitudinal surveys–comparing, combing, and interpreting using natural language processing. *Transportation Research Part C: Emerging Techniques*, 137, 103589.

Barnes, J., Morris, A., Welsh, R., Summerskill, S., Marshall, R., Kendrick, D., Logan, P., Drummond, A., Conroy, S., Fildes, B., Bell, J., 2016. Injuries to older users of buses in the UK. *Public Transport* 8(1), 25–38.

Bischof, J., Airoldi, E., 2012. Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK.

Björnstig, U., Bylund, P., Albertsson, P., Falkmer, T., Björnstig, J., Petzäll, J., 2005. Injury events among bus and coach occupants: non-crash injuries as important as crash injuries. *IATSS Research* 29(1), 79–87.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.

Bongini, P., Osborne, F., Pedrazzoli, A., Rossolini, M., 2022. A topic modelling analysis of white papers in security token offerings: which topic matters for funding? *Technological Forecasting and Social Change* 184, 122005.

Cambria, E., White, B., 2014. Jumping NLP curves: a review of natural language processing research. *IEEE Computational Intelligence Magazine* 9(2), 48–57.

Chang, F., Xu, P., Zhou, H., Lee, J., Huang, H., 2019. Identifying motorcycle high-risk traffic scenarios through interactive analysis of driver behavior and traffic characteristics. *Transportation Research Part F: Traffic Psychology & Behaviour* 62, 844–854.

Chen, Q., Chen, K., Ye, S., 2024. Noncollision injuries to passengers on buses: a case study from China. *Journal of Transport & Health* 35, 101776.

Chen, T., Lu, Y., Fu, X., Sze, N., Ding, H., 2022. A resampling approach to disaggregate analysis of bus-involved crashes using panel data with excessive zeros. *Accident Analysis & Prevention* 164, 106496.

Elawad, A., Murgovski, N., Jonasson, M., Sjöberg, J., 2024. Autonomous bus docking for optimal ride comfort of standing passengers. *IEEE Transactions on Intelligent Transportation Systems* 25(8), 9587–9596.

Elvik, R., 2019. Risk of non-collision injuries to public transport passengers: synthesis of evidence from eleven studies. *Journal of Transport & Health* 13, 128–136.

Goldberg, D.M., 2022. Characterizing accident narratives with word embeddings: improving accuracy, richness, and generalizability. *Journal of Safety Research*

80, 441–455.

Grimmer, J., Stewart, B.M., 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3), 267–297.

Halpern, P., Siebzehner, M.I., Aladgem, D., Sorkine, P., Bechar, R., 2005. Non-collision injuries in public buses: a national survey of a neglected problem. *Emergency Medicine Journal* 22(2), 108–110.

Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Techniques* 44, 363–381.

Hong Kong Transport Department, 2014. *Travel Characteristics Survey 2011*. https://www.td.gov.hk/filemanager/en/content_4652/tcs2011_eng.pdf.

Hong Kong Transport Department, 2024. *Road Traffic Accident Statistics 2023*. https://www.police.gov.hk/info/doc/statistics/traffic_report_2013_en.pdf.

Huang Z.R., Loo, B.P.Y., 2023. Urban traffic congestion in twelve large metropolitan cities: a thematic analysis of local news contents, 2009-2018. *International Journal of Sustainable Transportation* 17, 592–614.

Jing, P., Cai, Y., Wang, B., Wang, B., Huang, J., Jiang, C., Yang C., 2023. Listen to social media users: mining Chinese public perception of automated vehicles after crashes. *Transportation Research Part F: Psychology and Behaviour*, 93, 248–265.

Karekla, X., Fang, C., 2021. Upper body balancing mechanisms and their contribution to increasing bus passenger safety. *Safety Science* 133, 105014.

Karekla, X., Tyler, N., 2018. Reducing non-collision injuries aboard buses: passenger balance whilst walking on the lower deck. *Safety Science* 105, 128–133.

Karekla, X., Tyler, N., 2019. Reducing non-collision injuries aboard buses: passenger balance whilst climbing the stairs. *Safety Science* 112, 152–161.

Kendrick, D., Drummond, A., Logan, P., Barnes, J., Worthington, E., 2015. Systematic review of the epidemiology of non-collision injuries occurring to older people during use of public buses in high income countries. *Journal of Transport & Health* 2(3), 394–405.

Kuhn, K.D., 2018. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Techniques* 87, 105–122.

Kutela, B., Das, S., Dadashova, B., 2022a. Mining patterns of autonomous vehicle crashes involving vulnerable road users to understand the associated factors. *Accident Analysis & Prevention* 165, 106473.

Kutela, B., Langa, N., Mwende, S., Kidando, E., Kitali, A.E., Bansal, P., 2022b. A text mining approach to elicit public perception of bike-sharing systems. *Travel Behaviour and Society*, 24, 113–123.

Kwayu, K.M., Kwigizile, V., Lee, K., Oh, J.S., 2021. Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. *Accident Analysis & Prevention* 150, 105899.

Liu, C., Yang, S., 2022. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Systems with Applications* 207, 117991.

Loo, B.P.Y., Fan, Z., Lian, T., Zhang, F., 2023. Using computer vision and machine learning to identify bus safety risk factors. *Accident Analysis & Prevention* 185, 107017.

Lopez, D., Malloy, L.C., Arcoleo, K., 2022. Police narrative reports: do they provide end-users with the data they need to help prevent bicycle crashes? *Accident Analysis & Prevention* 164, 106475.

Lwanga, A., Mwanga, H.H., Mrema, E.J., 2022. Prevalence and risk factors for non-collision injuries among bus commuters in Dar es Salaam, Tanzania. *BMC Public Health* 22, 963.

Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272, Edinburgh, Scotland, UK.

Palacio, A., Tamburro, G., O'Neill, D., Simms, C.K., 2009. Non-collision injuries in urban buses—strategies for prevention. *Accident Analysis & Prevention* 41, 1–9.

Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2013. Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Techniques* 37, 177–192.

R Core Team, 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. 2004. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences* 101(9), 2658–2663.

Ramondt, S., Kerkhof, P., Merz, E., 2022. Blood donation narratives on social media: a topic modeling study. *Transfusion Medicine Reviews* 36(1), 58-65.

Ravenda, D., Valencia-Silva, M.M., Argiles-Bosch, J.M., García-Blandón, J., 2022. The strategic usage of Facebook by local governments: a structural topic modelling analysis. *Information & Management* 59, 103704.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G., 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4), 1064-1082.

Roberts, M.E., Stewart, B.M., Airoldi, E.M., 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association* 111(515), 988–1003.

Roberts, M.E., Stewart, B.M., Tingley, D., 2019. Stm: An R package for structural topic models. *Journal of Statistical Software* 91(1), 1–40.

Roque, C., Cardoso, J.L., Connell, T., Schermers, G., 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 & Prevention* 131, 336–349.

Rose, R.L., Puranik, T.G., Mavris, D.N., Rao, A.H., 2022. Application of structural topic modeling to aviation safety data. *Reliability Engineering and System Safety* 224, 108522.

Silvano, A.P., Ohlin, M., 2019. Non-collision incidents on buses due to acceleration and braking manoeuvers leading to falling events among standing passengers. *Journal of Transport & Health* 14, 100560.

Siman-Tov, M., Radomislensky, I., Marom, I., Kapra, O., Peleg, K., 2019. A nation-wide study on the prevalence of non-collision injuries occurring during use of public buses. *Journal of Transport & Health* 13, 164–169.

Taddy, M., 2013. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association* 108(503), 755–770.

Tao, S., Dai, T., Guo, Y., Wang, Y., Liu, B., Jiang, H., 2024. How do built environment characteristics influence bus use patterns across neighborhood types in

Beijing? A machine-learning analysis. *Travel Behaviour and Society* 35, 100756.

Tong, H.Y., Ng, K.W., 2021. A bottom-up clustering approach to identify bus driving patterns and to develop bus driving cycles for Hong Kong. *Environmental Science and Pollution Research* 28, 14343–14357.

Wali, B., Khattak, A.J., Ahmad, N., 2021. Injury severity analysis of pedestrian and bicyclist trespassing crashes at non-crossings: a hybrid predictive text analytics and heterogeneity-based statistical modeling approach. *Accident Analysis & Prevention* 150, 105835.

Wang, Y., Xiong, R., Yu, H., Bao, J., Yang, Z., 2022. A semantic embedding methodology for motor vehicle crash records: a case study of traffic safety in Manhattan Borough of New York City. *Journal of Transportation Safety & Security* 14(11), 1913–1933.

Xu, P., Xie, S., Dong, N., Wong, S.C., Huang, H., 2019. Rethinking safety in numbers: are intersections with more crossing pedestrians really safer? *Injury Prevention* 25(1), 20–25.

Xu, P., Zhou, H., Wong, S.C., 2021. On random-parameter count models for out-of-sample crash prediction: accounting for the variances of random-parameter distributions. *Accident Analysis & Prevention* 159, 106237.

Xu, P., Bai, L., Pei, X., Wong, S.C., Zhou, H., 2022. Uncertainty matters: Bayesian modeling of bicycle crashes with incomplete exposure data. *Accident Analysis & Prevention* 165, 106518.

Ye, Y., Zheng, P., Liang, H., Chen, X., Wong, S.C., Xu, P., 2024. Safety or efficiency? Estimating crossing motivations of intoxicated pedestrians by leveraging the inverse reinforcement learning. *Travel Behaviour and Society* 35, 100760.

Young, T., Hazarika, D., Poria, S., Cambria, E., 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13(3), 55–75.

Yue, L., Abdel-Aty, M., Wu, Y., Farid, A., 2020. The practical effectiveness of advanced driver assistance systems at different roadway facilities: system limitation, adoption, and usage. *IEEE Transactions on Intelligent Transportation Systems* 21(9), 3859–3870.

Zafari, B., Ekin, T., 2019. Topic modelling for medical prescription fraud and abuse detection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68(3), 751–769.

Zeng, Q., Wang, Q., Zhang, K., Wong, S.C., Xu, P., 2023. Analysis of the injury severity of motor vehicle–pedestrian crashes at urban intersections using spatiotemporal logistic regression models. *Accident Analysis & Prevention* 189, 107119.

Zhou, H., Yuan, C., Dong, N., Wong, S.C., Xu, P., 2020. Severity of passenger injuries on public buses: a comparative analysis of collision injuries and non-collision injuries. *Journal of Safety Research* 74, 55–69.

Zunjic, A., Sremcevic, V., Sijacki, V.Z., Sijacki, A., 2012. Research of injuries of passengers in city buses as a consequence of non-collision effects. *Work* 41(S1), 4943–4950.