From prediction to prevention: Using text mining and explainable machine learning for urban bus accident analytics[§]

Bowei Chen[†¶] Yufei Huang[‡] Yu Zheng[♯] Xiaofeng Liu[*]

[†]Adam Smith Business School, University of Glasgow, UK
[‡]Trinity Business School, Trinity College Dublin, Ireland
[♯]School of Finance, Southwestern University of Finance and Economics, China
[*]College of Artificial Intelligence and Automation, Hohai University, China

2024 Presentation @ Trinity College Dublin & University of Aberdeen

[§]Full Paper Publication @ *Risk Analysis*, 46, No. 1, e70183, 2026, doi.org/10.1111/risa.70183
[¶]✉ bowei.chen@glasgow.ac.uk

Road safety remains urgent global issue

The WHO Global Status Report on Road Safety 2023 indicates that since 2010, road traffic deaths have decreased by 5%, bringing the annual total to 1.19 million. However, road crashes remain a significant global health issue, with pedestrians, cyclists, and other vulnerable road users facing increasing and severe risks of death.

[1] https://www.who.int.

Bus accidents are important for several reasons

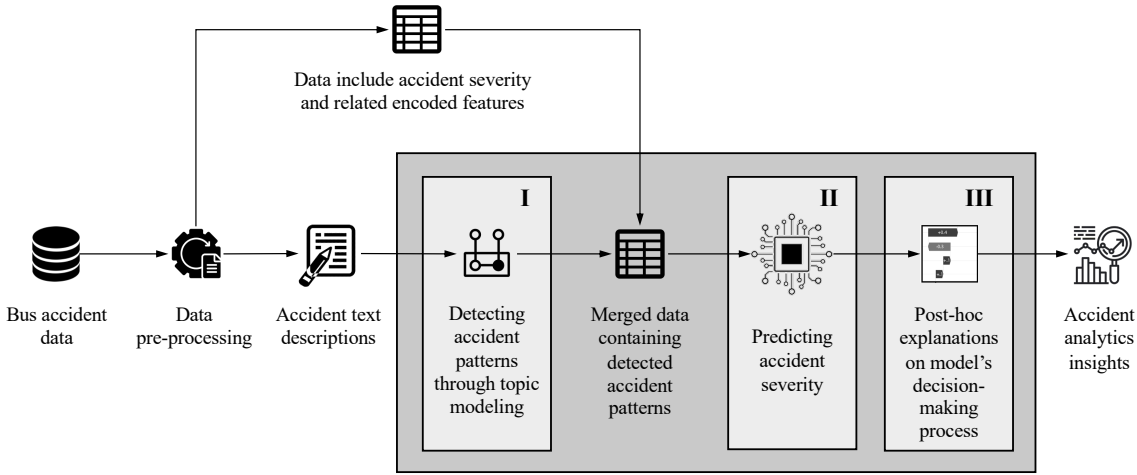
- **High casualty potential**: Buses often carry many passengers, so a single accident can cause significant injuries or fatalities, making each incident potentially very serious.
- **Public safety**: Buses are an essential part of public transport, particularly in densely populated areas. Ensuring their safety affects the whole community and helps maintain trust in transport systems.
- **Vulnerable populations**: Bus passengers often include school children, the elderly, and those on lower incomes, making their safety a key concern.[2]

[2] https://www.nhtsa.gov.

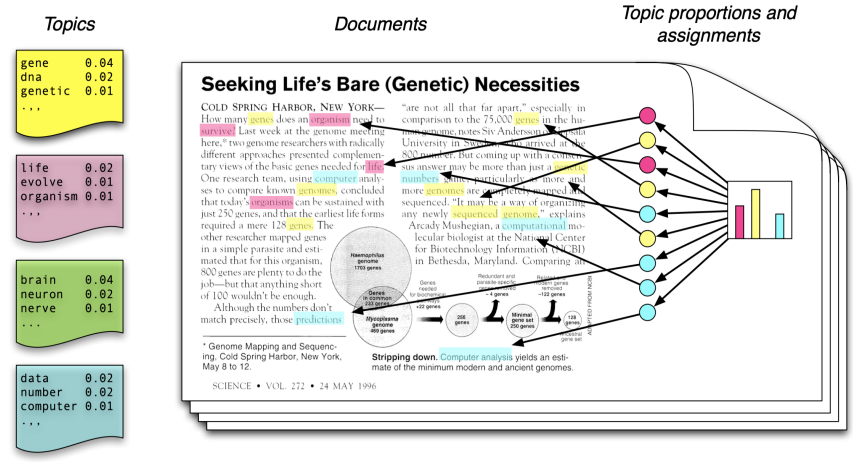
Data

- Our data was provided by transport companies from a tier-2 city in Jiangsu province, China, through a research collaboration. Commercial and personal information has been anonymised to comply with ethical agreements.
- The original Chinese text data was translated into English using the Google Cloud Translation API, following [1], and reviewed for accuracy.
- Standard pre-processing techniques were applied, including cleaning missing and inconsistent values.
- The final dataset includes 15,076 bus accidents from 243 routes across 36 companies, covering the period from 2013 to 2018.

Schematic view of the proposed analytical framework

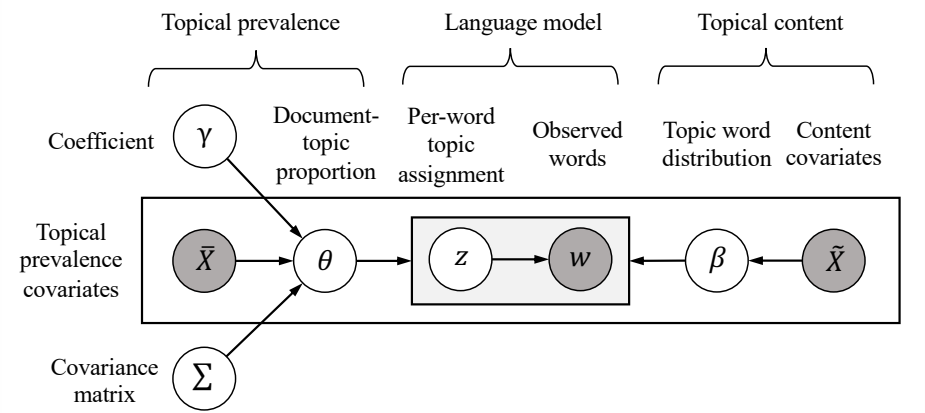


Stage I: Topic modeling[3]



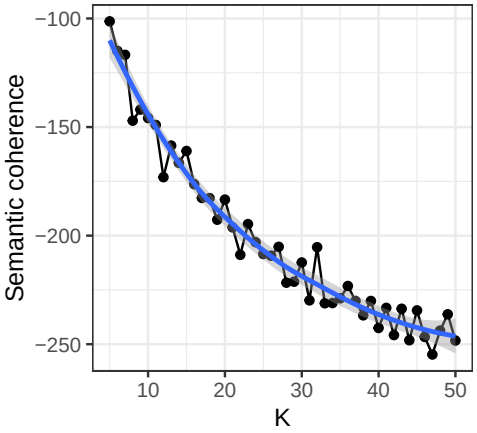
[3]Fig source: [2]

Structural topic model (STM) [3]



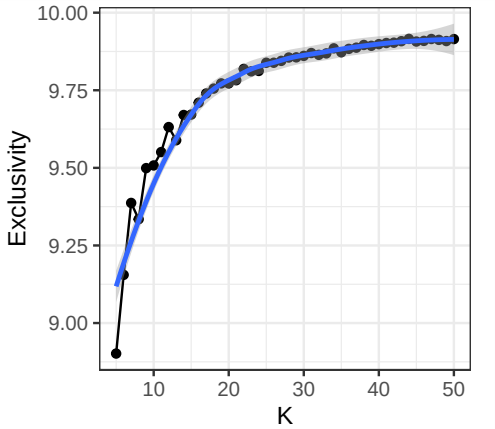
Determining the optimal number of patterns in the STM

- **Semantic coherence** [4] measures how frequently the most probable words in a topic co-occur in the same documents.
- High coherence means that the model has effectively grouped words that are commonly found together, which helps ensure that the topics are meaningful.



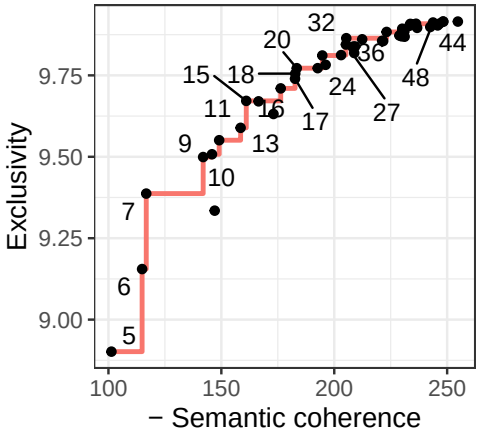
Determining the optimal number of patterns in the STM

- **Exclusivity** [5] measures how unique a topic's top words are, ensuring they do not appear in other topics.
- High exclusivity ensures that these top words are not shared by multiple topics, making the topics more distinct.



Determining the optimal number of patterns in the STM

- Combining these metrics enables fine-tuning of the STM to produce meaningful, non-overlapping topics.
- STM with $K = 7, 15, 20, 32$ offer the best trade-offs under a convex utility preference. Among these, $K = 15$ yields the most semantically interpretable results and is therefore used in our analysis.

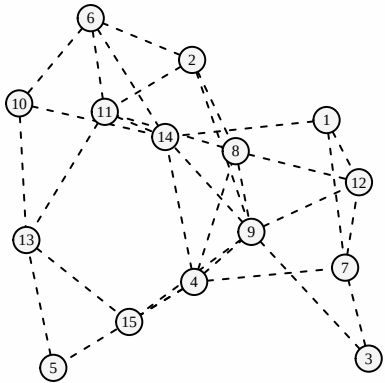


Detected bus accident patterns (topics)

Pattern	Description	Prop
1	Collision arises from bus reversing (with damage description)	5.94%
2	Rear-end collision between vehicles (at a detailed place)	7.54%
3	Sideswipe collision between vehicles (due to changing lanes)	18.92%
4	Rear-end collision (waiting for the green light, by cyclist)	4.16%
5	People injury accidents (with a detailed injury description, people sent to hospital)	4.32%
6	Sideswipe collision between vehicles (no detailed place and no damage description)	8.73%
7	Sideswipe collision between vehicles (no detailed place but with direction and damage description)	9.21%
8	Sideswipe collision between vehicles (at a detailed place, with direction but no damage description)	5.04%
9	Sideswipe collision between vehicles (left hand turns)	2.24%
10	Sideswipe collision with a fixed object (no detailed place and no damage description)	14.87%
11	Sideswipe collision between vehicles (at a detailed place, no direction and damage description)	3.60%
12	Sideswipe collision with a fixed object (at a detailed place, with damage description)	2.11%
13	No collision but people injuries with detailed description	7.86%
14	Sideswipe or rear-end collision (mainly due to road issues)	2.13%
15	Sideswipe or rear-end collision (no detailed place and no damage description)	3.32%

Correlation of bus accident patterns

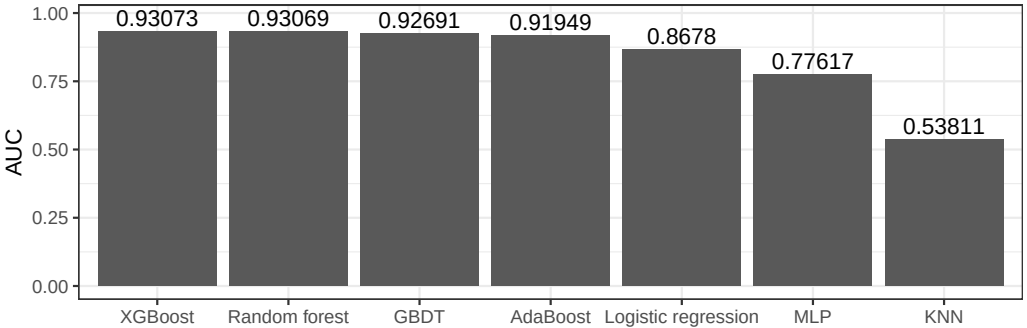
- Patterns 2, 6, 8, 10, 11 & 14 are about sideswipe or rear-end collisions but with fewer damage details;
- Patterns 1, 3, 7 & 12 provide damage details for different sideswipe or rear-end collisions;
- Patterns 5 & 13 describe bus accidents with people injuries.



Network visualization using the Meinshausen-Buhlmann method [6]

Stage II: Predictive models

- Input: 64 features after one-hot encoding, including variables from given data and topic modeling
- Output: 1 - severe accident; 0 - non-severe accident
- Model tuning using 10-fold cross-validation coupled with randomized search across hyper-parameter space.



Stage III: Post-hoc explanations

Post-hoc explanations are methods applied after a model makes predictions to explain how decisions were produced, particularly for complex "black-box" models like neural networks and ensemble methods.

Method	Pros	Cons
Impurity-based feature importance	Fast and efficient, built into tree-based models, and globally interpretable	Biased towards features with more categories, only applicable to tree-based models, and feature interaction ambiguity
Permutation feature importance	Model-agnostic and less biased	Can be computationally expensive, feature interaction ambiguity
Local interpretable model-agnostic explanations (LIME) [7]	Model-agnostic, provides local interpretability	Sensitivity to parameter settings, computationally expensive, and only local explanations
Shapley additive explanations (SHAP) [8]	Theoretically sound and consistent, handles interactions well	Computationally expensive, especially for large datasets

Shapley additive explanations (SHAP) [8]

- The general idea behind SHAP is to compute Shapley values from game theory to obtain both local and global insights into the contributions of feature values in the data.
- It can be computed as a weighted sum representing the marginal impact of each feature when added to the model, averaged over all possible feature combinations:

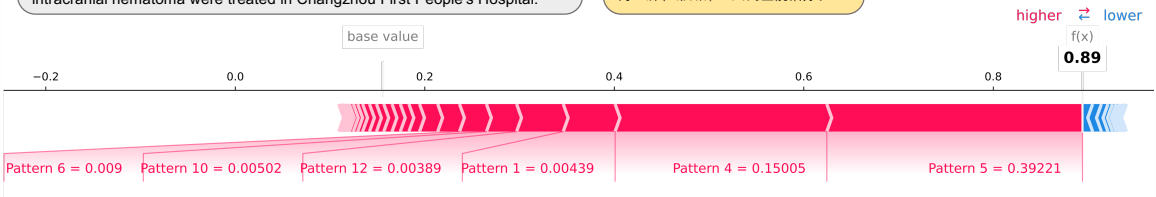
$$\xi_j(f,x) = \sum_{W \subseteq \{1,\dots,J\} \setminus \{j\}} \underbrace{\frac{|W|!(J-|W|-1)!}{J!}}_{=\textit{Weight}} \underbrace{\Big[f_x(W \cup \{j\}) - f_x(W) \Big]}_{=\textit{Contribution}},$$

where J is the size the feature vector, and W is the subsect of $\{1,\dots,J\} \setminus \{j\}$.

Examples of local explanations on feature importance

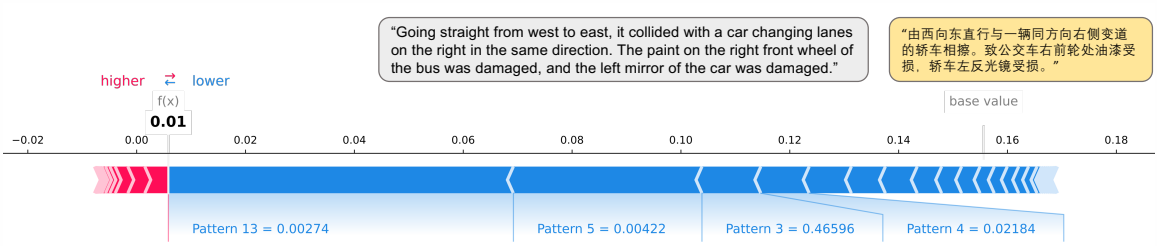
"Going straight from south to north, it collided with an electric scooter on the right in the same direction. Scratches on the rear panel of the bus and damage to the front panel of the electric scooter. The rider of the electric scooter, male 42-year-old Wang XXX, was injured. Injuries: Skull fracture and intracranial hematoma were treated in Changzhou First People's Hospital."

"由南向北直行与一辆右侧同方向电动车相撞。致公交车后傍板处擦伤，电动车前档板受损.电动车驾驶人 男 42岁 王XXX 受伤。伤情：颅骨骨折颅内血肿在常州第一人民医院治疗。"



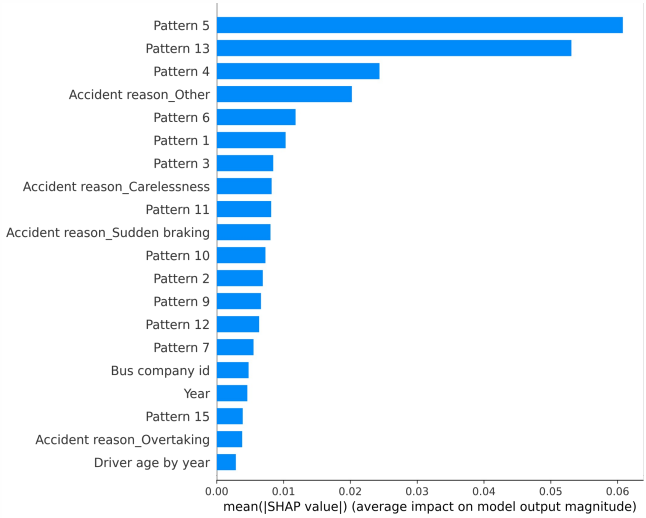
Pattern	Description
5	People injury accidents (with a detailed injury description, people sent to hospital)
4	Rear-end collision (waiting for the green light, by cyclist)
1	Collision arises from bus reversing (with damage description)
12	Sideswipe collision with a fixed object (at a detailed place, with damage description)
10	Sideswipe collision with a fixed object (no detailed place and no damage description)
6	Sideswipe collision between vehicles (no detailed place and no damage description)

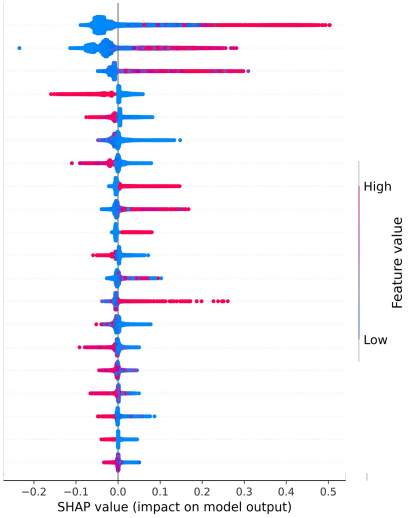
Examples of local explanations on feature importance



Pattern	Description
13	No collision but people injuries with detailed description
5	People injury accidents (with a detailed injury description, people sent to hospital)
3	Sideswipe collision between vehicles (due to changing lanes)
4	Rear-end collision (waiting for the green light, by cyclist)

Top 20 important global features (left) and their local explanations (right)





Insights & implications for bus drivers

- Carelessness, sudden braking, and underestimation are major causes of severe bus accidents.
- Non-injury collisions often occur due to overtaking or maintaining a short front distance.
- Bus drivers should be familiar with their route and timetable, and use technology to aid with gap decision, acceleration rate, and steering actions.
- Although seat-belts are not mandatory on city buses in many countries, drivers should ensure passengers are seated or holding a handrail before pulling away.
- Announcements like "Please hold on, the bus is about to move" should be maintained, particularly for elderly passengers, and additional warnings should be given in specific situations, such as deep downhill or gravel roads.

Insights & implications for bus companies

- Bus driver's age and years of driving experience are important factors in accident severity.
- Experienced young bus drivers are associated with a higher risk of severe accidents. On the one hand, younger drivers are more prone to severe accidents due to psychological characteristics such as impulsivity and risk-seeking behaviour. On the other hand, less experienced bus drivers may be more cautious and attentive, leading to fewer mistakes and a lower overall risk.
- Companies should focus on educating experienced young drivers and incentivizing safe behaviors to reduce the risk of severe accidents.

Insights & implications for government and public bodies

- Governments should prioritize the installation of dedicated left or right-turn lanes at intersections with high turn volumes or crash histories to improve safety.
- Public bodies should organize more promotional events to raise awareness of transport risks for passengers, pedestrians, and cyclists.
- Transport authorities should ensure that passengers are properly instructed to hold handrails and face forward while standing on buses to minimize the risk of injury.

Discussion

- The proposed framework integrates topic modeling, predictive modeling, and post-hoc explanations to analyze urban bus accident data – an approach not previously adopted in this literature (see paper Table 1).
- It is modular and flexible: Stage I supports other topic or transformer-based models; Stage II allows different predictive models; and Stage III accommodates various post-hoc explanation methods.
- Our model choice (STM+XGBoost+SHAP) balances predictive performance and interpretability: (i) STM incorporates document-level covariates and captures latent accident themes; (ii) XGBoost delivers strong predictive accuracy, handles nonlinearity and requires limited tuning; (iii) SHAP provides consistent, model-agnostic explanations at both local and global levels.

Thank you!

bowei.chen@glasgow.ac.uk
https://boweichen.github.io

Appendix: Evaluation metrics for topic models

- **Semantic coherence** [4] estimates the likelihood that an accident record contains the first few words of a pattern simultaneously. Given a list of the Q most probable words in pattern k, the semantic coherence for the pattern C_k is calculated as follows:

$$C_k = \sum_{e=2}^{Q} \sum_{\epsilon=1}^{e-1} \log \left\{ \frac{\Theta(v_e, v_\epsilon) + 1}{\Theta(v_\epsilon)} \right\},$$

where $\Theta(v_e, v_\epsilon)$ is the number of times words v_e and v_ϵ co-occur in the accident record text.

- **Exclusivity** [5] means the top words for a pattern are unlikely to appear within top words of other patterns, which can be computed using the frequency-exclusivity (FREX) score:

$$\mathrm{FREX}_{k,v} = \left[\frac{\pi}{\mathrm{ECDF}(\beta_{k,v}) / \sum_{\tilde{k}=1}^{k} \beta_{\tilde{k},v}} + \frac{1-\pi}{\mathrm{ECDF}(\beta_{k,v})} \right]^{-1},$$

where ECDF is the empirical cumulative distribution function and π is the weight.[4]

[4] In our used R package, $\pi = 0.7$ by default [9].

References

[1] C. Lucas, R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley, "Computer assisted text analysis for comparative politics," *Political Analysis*, vol. 23, pp. 254–277, 2015.

[2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[3] M. E. Roberts, B. M. Stewart, and E. M. Airoldi, "A model of text for experimentation in the social sciences," *Journal of the American Statistical Association*, vol. 111, no. 515, pp. 988–1003, 2016.

[4] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, vol. 1, pp. 262–272, 2011.

[5] E. M. Airoldi and J. M. Bischof, "Improving and evaluating topic models and other models of text," *Journal of the American Statistical Association*, vol. 111, pp. 1381–1412, 2016.

[6] T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman, "The huge package for high-dimensional undirected graph estimation in R," *Journal of Machine Learning Research*, vol. 13, pp. 1059–1062, 2012.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1135–1144, 2016.

[8] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 12, p. 4768–4777, 2017.

[9] M. E. Roberts, B. M. Stewart, and D. Tingley, "STM: An R package for structural topic models," *Journal of Statistical Software*, vol. 91, pp. 1–40, 2019.