Detecting Traffic Safety Risks using Cloud-sourced Data and Natural Language Model

Adewumi Augustine Adepitan, Adriana Valentina Farias, Shanjiang Zhu
aadepita@gmu.edu afarias4@gmu.edu szhu3@gmu.edu
Sid and Reva Dewberry Department of Civil, Environmental, and Infrastructure Engineering | George Mason University

Introduction

Active transportation has major health and environmental benefits, but pedestrians and cyclists still face serious safety. Crash records capture only some hazardous events and are reactively in nature. This study uses cloud-sourced near-miss reports and Natural Language Processing to reveal hidden risk patterns and identify high-risk locations for proactive safety action.

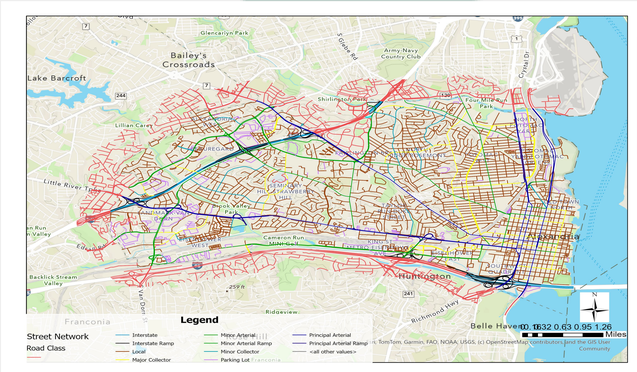

Figure1: Street Network by Road Class in the Study Area (Alexandria, VA)

Data

- Near-miss reports: 3,000+ cloud-sourced incidents from Northern Virginia Families for Safe Streets (NoVA FSS) (pedestrian/bicyclist safety narratives).
- Crash records: 5,000+ VDOT police-reported pedestrian/bicyclist crashes from 2017-2025 in the same study area.

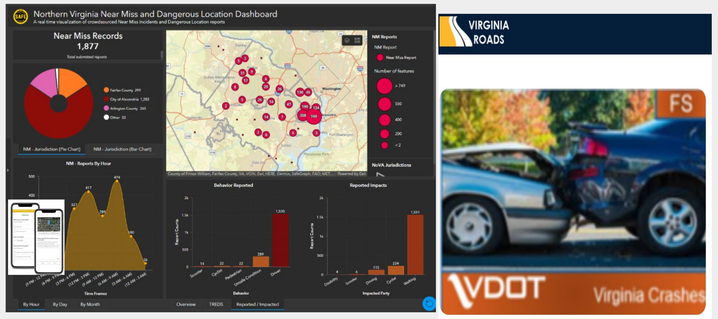

Figure 2: Data source (Near mises and Crash data)

Research Objectives

- Develop an approach that automatically processes unstructured textual data for safety applications;
- Identify prevalent themes and risk factors contributing to pedestrian and cyclist safety based on actual crash reports and self-reported near-miss data;
- To provide a scalable methodology that can be extended to other datasets and regions.

Methodology

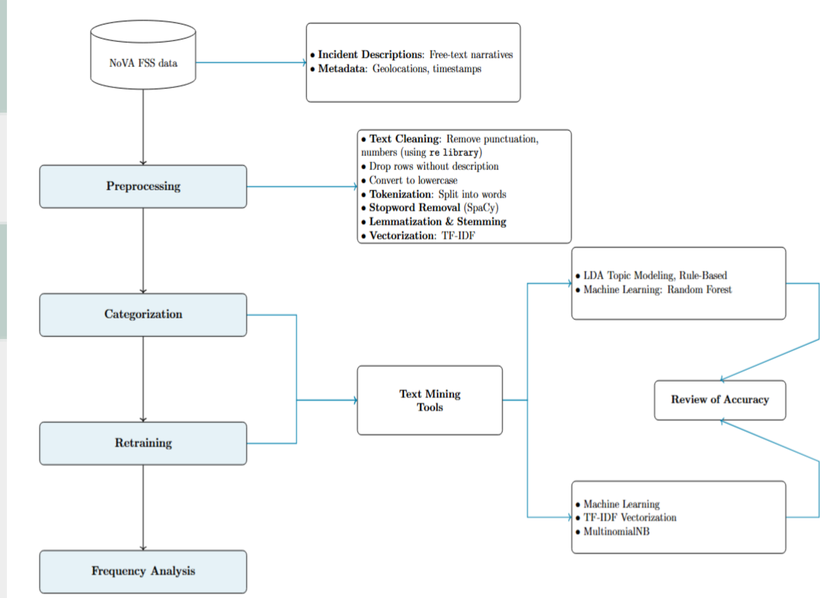

Figure 3: Flow Chart of the proposed NLM Method

Incident Category	Frequency	Percentage
Running Red light or Stop Signs	870	37.04
Crosswalk Issues	621	26.44
Parking in Sidewalk/Bike Lane	592	25.20
Drivers Not Yielding	139	5.92
Speeding or Reckless Driving	103	4.38
Scooter Noiseless Hazard	11	0.47
Dangerous Intersection	9	0.38
No sidewalk	4	0.17
Grand Total	2,349	100

Table 1: Frequency and Percentage of Incident Categories Reported in Near-Miss Dataset

Zero-Inflated Poisson (ZIP) model (two-part):

1) Inflation (logit): probability an intersection is "always-zero"

$$p_i = \frac{\exp(\gamma_0 + \gamma^T Z_i)}{1 + \exp(\gamma_0 + \gamma^T Z_i)}$$

2) Count (Poisson): crash rate when crashes occur

$$\lambda_i = \exp(\beta_0 + \beta^T X_i)$$

$Y_i \sim \text{Poisson}(\lambda_i)$ (conditional on not being always-zero)

Mixture probabilities:

$$P(Y_i = 0) = p_i + (1 - p_i)\exp(-\lambda_i)$$

$$P(Y_i = y) = (1 - p_i)\frac{\lambda_i^y \exp(-\lambda_i)}{y!}, \quad y = 1, 2, \ldots$$

Notation (poster-friendly):

Y_i = crash count at intersection i; p_i = probability of being "always-zero"; λ_i = expected crashes if not always-zero.
Z_i = road context (e.g., functional class, speed category); X_i = NLP-derived near-miss indicators; β, γ = coefficients.

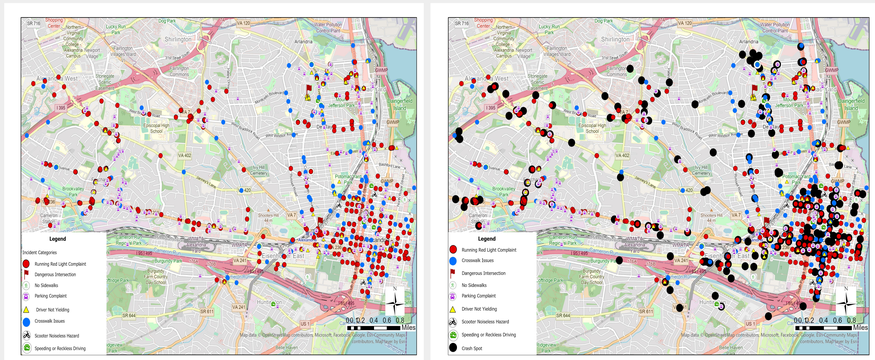

Figure 4: Spatial Distribution and correlation of near-miss occurrence with crash Spot

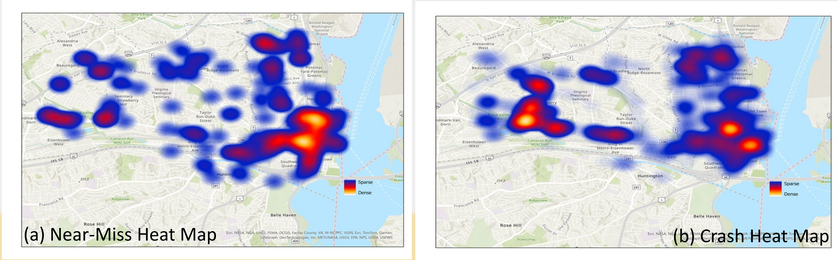

Figure 5: Near-Miss vs Crash Heat Maps

Analysis Results

Zero-inflated Poisson regression
Inflation model: logit

Number of obs = 2,106
Nonzero obs = 158
Zero obs = 1,948
LR chi2(3) = 84.93
Prob > chi2 = 0.0000

Log likelihood = -602.5952

Crash_Count	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
RuningRedlight	1.248728	0.1750678	7.13	0.000	0.9056011	1.591854
crosswalkIssues	0.6062113	0.1958521	3.10	0.002	0.2223484	0.9900743
driverNotyield	0.5356038	0.2500605	2.14	0.032	0.0454942	1.025713
_cons	-1.928749	0.1851186	-10.42	0.000	-2.291575	-1.565924
inflate: func_class						
Collector	-1.525333	0.4033908	-3.78	0.000	-2.315964	-0.7347011
Major Arterial	-2.372457	0.6455501	-3.68	0.000	-3.637712	-1.107202
Minor Arterial	-1.665156	0.4902326	-3.40	0.001	-2.625994	-0.704318
_cons	1.202742	0.2717198	4.43	0.000	0.6701812	1.735303

TABLE 2: Zero-inflated Poisson regression (ZIP) model results with Road class as inflation covariate

- Near-misses as crash surrogates (cloud-sourced + NLP → ZIP model). Using cloud-sourced near-miss reports, we applied NLP to extract key risk behaviors from natural language narratives by road users (e.g., red-light running, crosswalk conflicts, failure-to-yield) and aggregated them to the intersection level, then linked them with observed crash counts.
- The ZIP results show these NLP-derived near-miss signals are significantly associated with higher crash frequency, strongest for red-light running, followed by crosswalk issues and failure-to-yield ($p < 0.05$).
- This demonstrates that near-miss data can act as an early warning system, revealing crash-prone locations before sufficient crash history accumulates and helping prioritize targeted countermeasures (signal timing/enforcement, crosswalk upgrades, yielding treatments), especially on collector and arterial corridors.

ACKNOWLEDGEMENTS

This study is funded by the National Science Foundation grant #2431423: Advancing Community-Centered Solutions for Walking, Biking, and Micromobility. We would also like to thank VDOT and NoVA FSS for supporting our Study. All the All statements are the authors' and do not reflect the views of any government agency.