# Insurance Data Science 2025: Application of the NLP models in loss modeling for actuarial science

Manuel Caccone

June 19, 2025

# Preface

# The Traditional Approach Problem

## GLM Framework Limitations[1]

- **Standard models**: $N_j \sim Poisson(\lambda_j)$, $Y_j \sim Gamma(\alpha_j, \beta_j)$[2]
- **Insufficient personalization** - coefficients miss deeper patterns
- **Limited flexibility** with policyholder dependence[3]
- Classic clustering (*Chi-squared*, *K-means*[4]) cannot capture complexity

## Core Problems

- **Misclassification** of loss events[5]
- **High noise** from randomness and lack of information
- Missing **"cross-existence" risks** between policyholders

[1] Wüthrich and Buser (2023); Goldburd, Khare, and Tevet (2016); Ohlsson and Johansson (2010)

[2] Frees (2008); Antonio and Verbelen (2023)

[3] Frees (2008); Antonio and Verbelen (2023)

# The NLP Solution

**The Misclassification Mathematical Problem**

$\mu_{i,j}^F = \mathbb{E}[X_{i,j}^F] \neq \mu_{i,j}^T = \mathbb{E}[X_{i,j}^T]$

Where $X_{i,j}^F$ = misspecified peril, $X_{i,j}^T$ = true peril

**NLP Advantage**[6]

- **"Pre-clustering via NLP"** prevents misclassification
- Extract **semantic context** from claims text
- Capture **hidden risk factors** beyond structured variables
- Enable **context-aware clustering** for better risk profiling

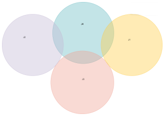

Figure 1: Eulero-Venn coefficients context

# How to?

# Collecting the data

- First, we need to collect the data, which can include the following:
  - the **policyholder's declaration**;
  - the **loss adjuster's evaluation**;
  - the **loss data**.

Figure 2: Loss Documents

# Introducing NLP in Actuarial Analysis

## From Structured to Semantic Analysis

- **Classic limitations**: Noise, imprecise classification, missing textual context
- **NLP breakthrough**: Extract insights from claim/crash descriptions[7]

## Text Embeddings Advantage

- Capture **semantic meaning** and contextual relationships and precise risk profiling

## Domain-Specific Fine-tuning Challenge

- "Generalist" models miss **insurance technical language** → **Solution**: Fine-tuned GPT2-Small on synthetic insurance Q&A pairs
- **Result**: Insurance-optimized embeddings for actuarial analysis

[7] Devlin et al. (2018); Xu, Manathunga, and Wei (2022); Zappa, Borrelli, et al. (2021)

# Topic Modeling & BERTopic Framework

## BERTopic: 4-Stage Process[8]

1. **Embedding Generation**: Text → numerical vectors
2. **Dimensionality Reduction**: UMAP complexity reduction[9]
3. **Clustering**: HDBSCAN groups similar embeddings[10]
4. **Topic Representation**: Extract key descriptive words

## Actuarial Value

- Discover **recurring patterns** in large document collections
- Uncover **hidden risk factors** not apparent from structured variables
- Reveal **typical incident scenarios** for risk quantification

[8] Grootendorst (2022)

[9] McInnes, Healy, and Melville (2018)

[10] McInnes, Healy, and Astels (2017)

# BERTopic: A Powerful Approach for Large Text Volumes

- BERTopic is particularly well-suited for **large datasets** due to its ability to use GPU-accelerated implementations (cuML for UMAP and HDBSCAN), providing a 10-50x speedup[11].


Figure 3: BERTopic

[11] Allaoui, Kherfi, and Cheriet (2020); McInnes, Healy, and Melville (2018)

# Application

# Applying BERTopic to Crash Data (NMVCCS)[12]

### Automated Pattern Discovery

- Applied to NMVCCS textual crash descriptions and discovered **semantic patterns (topics)** automatically

### Key Pattern Examples

- Standard two-vehicle accidents (-1)
- Pre-crash critical events (0)
- Intersection left-turn collisions (2)
- Safety-mitigated events with seatbelts (3)

### Actuarial Intelligence

- Transform semantic patterns into **risk profiles**
- **Intersection left-turn crashes**: Highest risk (5.88% mortality)
- **Pre-crash critical events**: Medium-high injury risk

[12] National Highway Traffic Safety Administration (2008); National Highway Traffic Safety Administration (2007)

# From BERTopic Topics to Actuarial Risk

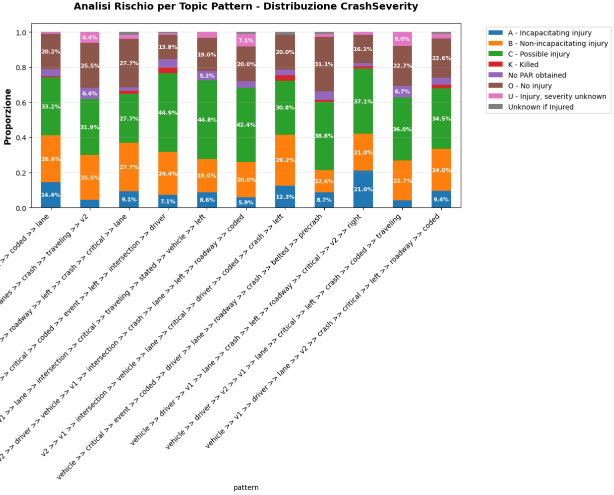


Figure 4: Pattern distribution

# From BERTopic Topics to Actuarial Risk

**High-Risk Patterns Identified:**

- "Vehicle → Driver → Event → Coded" shows highest fatality rate (20.2%)
- Intersection-related patterns consistently show elevated injury severity
- Most patterns dominated by "possible injury" and "no injury" outcomes

**Key Observations:**

- Fatal crashes represent 5-10% across most patterns
- Incapacitating injuries are consistently the smallest category
- Pattern complexity suggests sophisticated crash sequence analysis

# From BERTopic Topics to Actuarial Risk

**Data Considerations:**

- Pattern distribution shows balanced representation across scenarios

**Strategic Applications:**

- Use high-severity patterns for targeted underwriting
- Leverage pattern-specific data for actuarial modeling

ISOA

# Demographic Risk Profiling with Topic Insights

## Key Findings from 1,586 Records

- High-risk groups: **Males 36-45** and **Males 65+** (Risk Score 1.79)
- Reveals **"Volume vs. Risk Paradox"** - highest risk highest volume

## Gender-Specific Patterns

- **Males**: Higher crash frequency
- **Females**: Experience higher injury severity in comparable crashes

## Actionable Insights

- Male risk pattern: Intersection Complexity (Risk Score 2.15)
- Female risk pattern: Vehicle-Driver Critical (Risk Score 2.42)

# Demographic Risk Profiling with Topic Insights

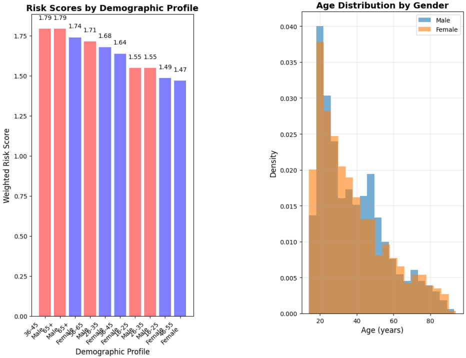


Figure 5: Demographic considerations

# Demographic Risk Profiling with Topic Insights

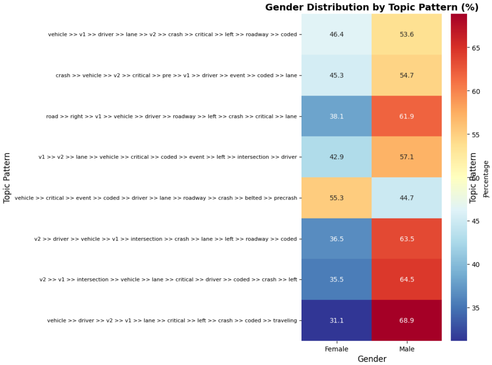


Figure 6: Demographic considerations

# Demographic Risk Profiling with Topic Insights

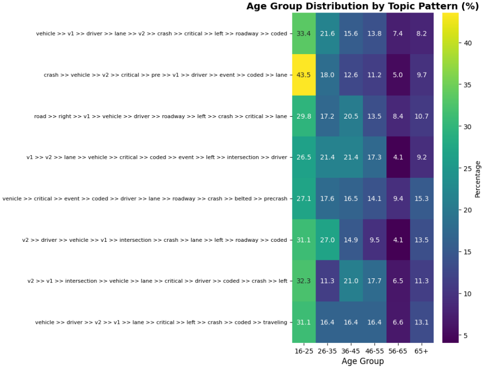


Figure 7: Demographic considerations

# Demographic Risk Profiling with Topic Insights

**Gender Distribution Insights:**

- Most patterns show 60-70% male involvement, confirming higher male crash frequency
- "Vehicle → Driver → V2 → V1 → Lane → Critical" shows highest male concentration (68.9%)
- Age distribution varies significantly by pattern - some skew younger (16-25), others toward middle age (36-45)

**Pattern-Specific Demographics:**

- Complex intersection patterns tend to involve older drivers (46-65+)
- Simple lane-change patterns show higher younger driver involvement
- Critical/traveling patterns demonstrate mixed age distributions

# Demographic Risk Profiling with Topic Insights

**High-Risk Demographics Identified:**

- Males 36-45 and Males 65+ both score 1.79 (highest risk categories)
- Females consistently show lower risk scores across age groups
- Risk scores range from ~1.47 to 1.79, indicating meaningful differentiation

**Actuarial Applications:**

- **Volume vs. Risk Paradox**: High-risk groups aren't always highest volume
- Gender-specific pattern targeting needed (males: frequency, females: severity)
- Age-based risk profiling shows clear segmentation opportunities for pricing

ISOA

# Dashboard

# Exploring Insights: The Interactive Dashboard

- We have translated complex data and models into **actionable actuarial insights**.
- These results can be explored interactively through our dedicated **Interactive Live Results Dashboard** .
- It offers key visualizations such as **Demographic Risk Profiling** and **BERTopic Topic Modeling Results**
- Gain deeper understanding of **crash-patterns** based **3D reconstruction of the types of accident**.
- Features include **Risk Score Heatmaps**, **Interactive Topic Clustering**

# Exploring Insights: The Interactive Dashboard

- **Dashboard:** Compatible with modern browsers (Chrome, Firefox, Safari, Edge) and based on NMVCCS crash data and insurance claims analysis.

Launch the Interactive Dashboard to explore the data:

bitly

ISOA

# Replicate It!

# Explore my code

- **github.com/manuelcaccone/NLP-Actuarial-Loss-Modeling:** Compatible with modern browsers (Chrome, Firefox, Safari, Edge) and based on NMVCCS crash data and insurance claims analysis.

Launch the Interactive Dashboard to explore the data:

bitly

**Visit the GitHub repository to view the source code and contribute**

ISOA

# Conclusion

# Benefits of NLP-Based Approach for Actuaries

### Context Enhancement

- Extract deep insights from unstructured text beyond structured variables

### Smart Clustering

- Group claims/policyholders by semantic patterns, not just demographics

### Risk Quantification

- Link specific incident scenarios to measurable risk profiles (severity, mortality)

### Fraud Detection

- Identify suspicious linguistic patterns and potential misclassifications[13]

[13] Gomes, Sousa, and Lopes (2021); Contributors (2023)

# Thank you

Manuel Caccone
**AI Task Force**
Italian Society of Actuaries
manuel.caccone@gmail.com

# References I

Allaoui, Mebarka, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. 2020. "Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study." *Lecture Notes in Computer Science* 12119: 317–25.

Antonio, Katrien, and Roel Verbelen. 2023. "Claim Frequency Modeling in Insurance Pricing Using GLM, Deep Learning, and Gradient Boosting." 2023. https://aktuar.de/en/knowledge/specialist-information/detail/claim-frequency-modeling-in-insurance-pricing-using-glm-deep-learning-and-gradient-boosting/.

# References II

Artís, Manuel, Mercedes Ayuso, and Montserrat Guillén. 2002. "Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims." *Journal of Risk and Insurance* 69 (3): 325–40.

Boulieris, Petros, John Pavlopoulos, Alexandros Xenos, and Vasilis Vassalos. 2023. "Fraud Detection with Natural Language Processing." *Machine Learning*, 1–22.

Contributors, ResearchGate. 2023. "Survey on Insurance Claim Analysis Using Natural Language Processing and Machine Learning." *ResearchGate*.

# References III

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv Preprint arXiv:1810.04805*.

Frees, Edward W. 2008. *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Goldburd, Mark, Anand Khare, and Dan Tevet. 2016. "Generalized Linear Models for Insurance Rating." *Casualty Actuarial Society E-Forum*.

Gomes, Susana, João Sousa, and Joaquim Lopes. 2021. "Insurance Fraud Detection with Unsupervised Deep Learning." *Journal of Risk and Insurance* 88 (3): 591–618.

# References IV

Grootendorst, Maarten. 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." https://github.com/MaartenGr/BERTopic.

McInnes, Leland, John Healy, and Steve Astels. 2017. "Hdbscan: Hierarchical Density Based Clustering." *Journal of Open Source Software* 2 (11): 205.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv Preprint arXiv:1802.03426*.

National Highway Traffic Safety Administration. 2007. "National Motor Vehicle Crash Causation Survey Dataset." U.S. Department of Transportation.

ISOA

# References V

———. 2008. "National Motor Vehicle Crash Causation Survey: Report to Congress." DOT HS 811 052. U.S. Department of Transportation.

Ohlsson, Esbjörn, and Björn Johansson. 2010. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer.

Pitkänen, Paavo. 1975. "Tariff Theory." *ASTIN Bulletin* 8 (2): 204–28. https://doi.org/10.1017/S0515036100009338.

Vandervorst, Félix, Wouter Verbeke, and Tim Verdonck. 2022. "Data Misrepresentation Detection for Insurance Underwriting Fraud Prevention." *Decision Support Systems* 161: 113690.

# References VI

Wüthrich, Mario V., and Christoph Buser. 2023. "Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting." *Risks* 11 (9): 163.

Xu, Shuzhe, Vajira Manathunga, and Libo Wei. 2022. "Framework of BERT-Based NLP Models for Frequency and Severity in Insurance Claims." *Variance* 15 (2).

Zappa, Diego, Mattia Borrelli, et al. 2021. "Text Mining in Insurance: From Unstructured Data to Meaning." *Variance*.

ISOA