european
actuarial
academy

The European knowledge center for actuaries

AI’s Transformative Power: NLP for Next-Generation Actuarial Risk Assessment

Manuel Caccone

4 December 2025

Manuel Caccone

ABOUT ME

Manuel Caccone

Italian Society of Actuaries AI TF

- **10+ years of experience** in insurance and actuarial sector, Full Stack Actuary
- Gruppo Unipol: Life & Non-Life Risk Manager, specialized in Internal Model Premium Risk, Catastrophe Modeling, ORSA, ESG Risk, Cyber Risk
- **Expert in Legal Tech** - developing AI solutions (autonomous agents, RAG systems) for legal applications
- **IAA AI Task Force Leader** - guiding strategic initiatives at the intersection of AI and actuarial science
- Senior Actuary with expertise in AI, Machine Learning, Big Data, Software Developing

Constraints of GLM-Based Modeling[1]

- **Conventional approaches**: $N_j \sim Poisson(\lambda_j)$, $Y_j \sim Gamma(\alpha_j, \beta_j)$[2]
- **Inadequate customization** - model parameters overlook nuanced relationships
- **Restricted adaptability** when modeling policyholder interdependence[3]
- Traditional segmentation methods (*Chi-squared*, *K-means*[4]) fail to handle intricacy

---

[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)

[4] Pitkänen (1975)

Fundamental Issues

- **Incorrect categorization** of claim incidents[5]
- **Excessive variability** due to stochastic nature and data gaps
- Overlooking **“interrelated hazards”** among insured parties

[5] Vandervorst, Verbeke, and Verdonck (2022); Artís, Ayuso, and Guillén (2002)

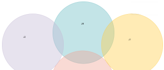
Mathematical Framework of the Categorization Issue

$$\mu^F_{i,j} = \mathbb{E}[X^F_{i,j}] \neq \mu^T_{i,j} = \mathbb{E}[X^T_{i,j}]$$

Here $X^F_{i,j}$ = incorrectly identified hazard, $X^T_{i,j}$ = actual hazard

Benefits of NLP Implementation[6]

- **"Preliminary NLP-based segmentation"** eliminates categorization errors
- Derive **meaning-based understanding** from textual claim data
- Identify **latent exposure elements** not visible in traditional data fields
- Support **meaning-driven grouping** for improved exposure assessment



- Initially, information compilation is required, potentially encompassing:
  - **statements from the insured party**;
  - **assessment by the claims examiner**;
  - **claim-related statistics**.

Figure: Loss Documents

Evolution from Traditional to Meaning-Based Evaluation

- **Traditional constraints**: Variability, inaccurate categorization, absent narrative information
- **NLP innovation**: Derive understanding from incident/accident narratives[7]

Benefits of Textual Representations

- Encode **linguistic significance** plus relational context and accurate exposure classification

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

Sector-Specific Adaptation Requirements

- “Universal” algorithms overlook **sector-specific terminology** → **Approach**: Adapted GPT2-Small using artificial insurance question-response data
- **Outcome**: Sector-enhanced representations for insurance analytics

BERTopic: Quadruple-Phase Methodology[8]

1. **Vector Creation**: Narrative → mathematical representations
2. **Space Compression**: UMAP dimension simplification[9]
3. **Grouping**: HDBSCAN identifies related vectors[10]
4. **Theme Identification**: Derive characteristic terminology

[8] Grootendorst (2022)
[9] McInnes, Healy, and Melville (2018)
[10] McInnes, Healy, and Astels (2017)

Insurance Application

- Identify **repeating structures** within extensive text repositories
- Expose **concealed exposure elements** beyond traditional data points
- Uncover **common occurrence patterns** for exposure assessment

BERTopic: An Effective Solution for Extensive Textual Collections

- BERTopic excels with **extensive corpora** through its capability for GPU-enhanced processing (cuML implementations of UMAP and HDBSCAN), delivering 10-50x performance improvements[11].

Figure: BERTopic

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

Implementing BERTopic on Collision Records (NMVCCS)[12]

Pattern Recognition Automation

- Utilized on NMVCCS narrative accident reports and identified **meaning-based categories (themes)** without manual intervention

Notable Theme Instances

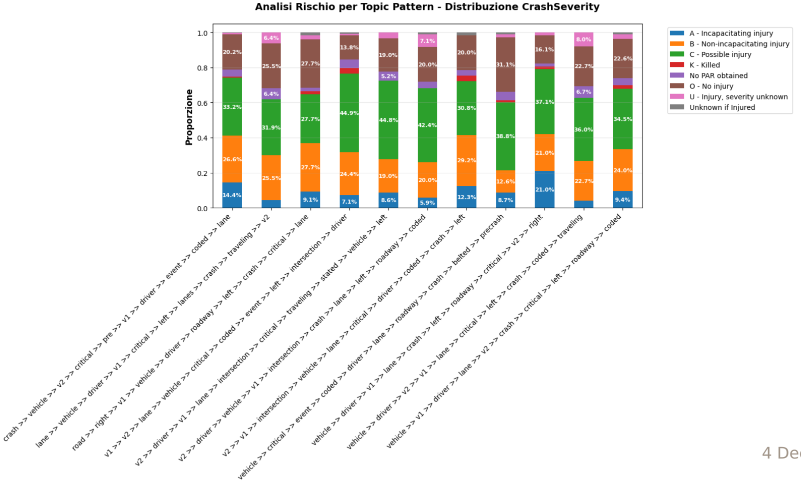
- Regular dual-vehicle incidents (-1)
- Events preceding collision occurrence (0)
- Junction-based turning incidents (2)
- Protection-enhanced situations involving restraints (3)

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

Insurance Analytics Value

- Convert meaning-based categories into **exposure classifications**
- **Junction turning incidents**: Maximum exposure (5.88% fatality rate)
- **Events preceding collision**: Moderate to elevated harm exposure

Converting BERTopic Themes into Insurance Exposure



Converting BERTopic Themes into Insurance Exposure

Elevated-Exposure Categories Discovered:

- "Vehicle → Driver → Event → Coded" exhibits maximum lethality percentage (20.2%)
- Junction-associated categories regularly demonstrate increased harm intensity
- Majority of categories show prevalence of "potential harm" and "unharmed" results

Converting BERTopic Themes into Insurance Exposure

Principal Insights:

- Lethal incidents comprise 5-10% within majority of categories
- Severe disabilities remain the least frequent outcome
- Category intricacy indicates advanced incident progression examination

Converting BERTopic Themes into Insurance Exposure

Information Analysis:

- Category spread demonstrates equal coverage of different situations

Business Implementation:

- Apply elevated-intensity categories for focused risk assessment
- Utilize category-targeted information for insurance calculations

Primary Discoveries from 1,586 Entries

- Elevated-exposure segments: **Men 36-45** and **Men 65+** (Exposure Index 1.79)
- Uncovers **“Quantity versus Exposure Contradiction”** - maximum exposure ≠ maximum quantity

Sex-Based Characteristics

- **Men**: Elevated incident occurrence
- **Women**: Demonstrate increased harm intensity in similar incidents

Practical Applications

- Male exposure characteristic: Junction Intricacy (Exposure Index 2.15)
- Female exposure characteristic: Vehicle-Operator Essential (Exposure Index 2.42)

Population-Based Exposure Assessment Using Theme Analysis

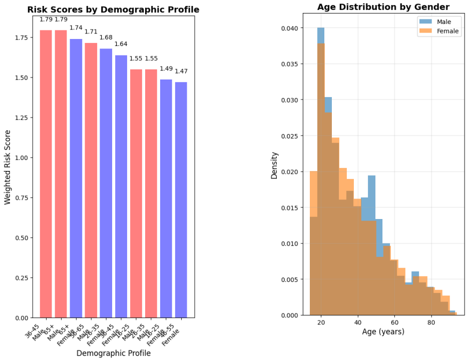


Figure: Demographic considerations

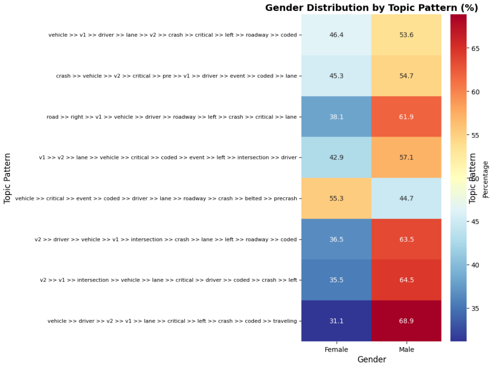


Figure: Demographic considerations

Population-Based Exposure Assessment Using Theme Analysis

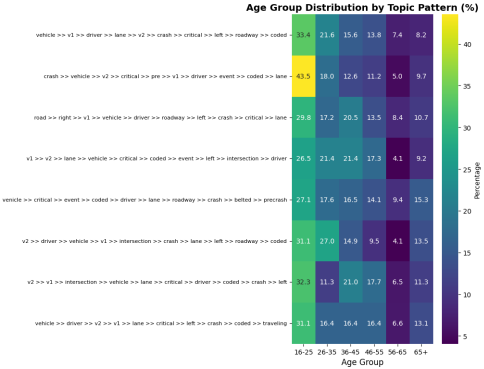


Figure: Demographic considerations

Sex Allocation Observations:

- Majority of categories indicate 60-70% male participation, validating elevated male incident occurrence
- “Vehicle → Driver → V2 → V1 → Lane → Critical” demonstrates maximum male representation (68.9%)
- Age spread differs considerably across categories - certain categories favor youth (16-25), while others favor middle-aged individuals (36-45)

Population-Based Exposure Assessment Using Theme Analysis

Category-Related Population Characteristics:

- Intricate junction categories frequently include senior operators (46-65+)
- Basic lane-shift categories demonstrate elevated young operator participation
- Essential/movement categories show diverse age spreads

Elevated-Exposure Populations Discovered:

- Men 36-45 and Men 65+ each register 1.79 (maximum exposure groups)
- Women regularly demonstrate reduced exposure indices throughout age ranges
- Exposure indices vary from approximately 1.47 to 1.79, showing substantial distinction

Insurance Implementation:

- **Quantity versus Exposure Contradiction**: Elevated-exposure segments don’t always represent maximum quantity
- Sex-based category focus required (men: occurrence, women: intensity)
- Age-oriented exposure assessment reveals distinct opportunities for rate determination

Discovering Patterns: The Dynamic Interface

- Complex information and algorithms have been converted into **practical insurance intelligence**.
- These findings are accessible through our specialized **Dynamic Analysis Interface**.
- The platform provides essential displays including **Population Exposure Assessment** and **BERTopic Theme Discovery Outputs**
- Develop comprehensive knowledge of **incident structures** through **spatial visualization of collision categories**.
- Capabilities encompass **Exposure Index Matrices**, **Dynamic Theme Groupings**

Discovering Patterns: The Dynamic Interface

- **Interface:** Functions across current web platforms (Chrome, Firefox, Safari, Edge) utilizing NMVCCS collision records and claim investigation data.

Access the Dynamic Interface for data investigation:

bitly

Access the implementation

- **github.com/manuelcaccone/NLP-Actuarial-Loss-Modeling:** Functions across current web platforms (Chrome, Firefox, Safari, Edge) utilizing NMVCCS collision records and claim investigation data.

Access the Dynamic Interface for data investigation:

bitly

Access the GitHub repository for implementation details and collaboration

Contextual Enrichment: Derive comprehensive understanding from narrative information exceeding traditional data fields
Intelligent Segmentation: Organize incidents/insured parties through meaning-based structures, beyond simple population metrics
Exposure Assessment: Connect particular occurrence patterns with quantifiable exposure characteristics (intensity, lethality)
Anomaly Identification: Detect irregular textual structures and classification errors[13]

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

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

References

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.

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*.

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.

References

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.
Grootendorst, Maarten. 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure."
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.
National Highway Traffic Safety Administration. 2008. "National Motor Vehicle Crash Causation Survey: Report to Congress." DOT HS 811 052. U.S. Department of Transportation.

References

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