XL Insurance

Tracking Trending Topics in Insurance for Emerging Risk Identification

Amin Karbassi, Ph.D. | Senior Casualty Risk Consultant

Insurance Data Science Conference - London | June 2025

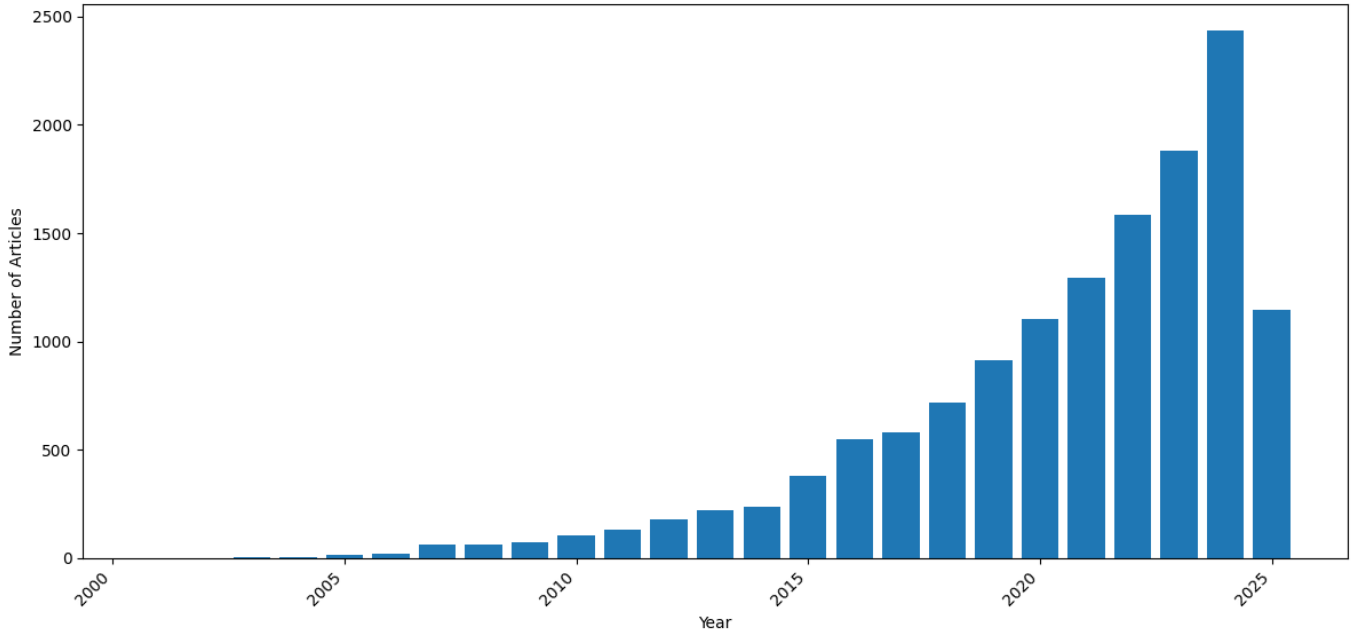
Emerging risks

- Newly developing or evolving risks that are not fully understood or quantified
- Often first appear in scientific literature before public recognition
  - PFAS (1990s and early 2000s), Asbestos (1960s), and Opioid Crisis (mid-1990s)
- Traditional clustering methods often ignore time element, limiting trend identification

Goal: Identify **trends** that are not only **thematically coherent** but also **temporally focused**

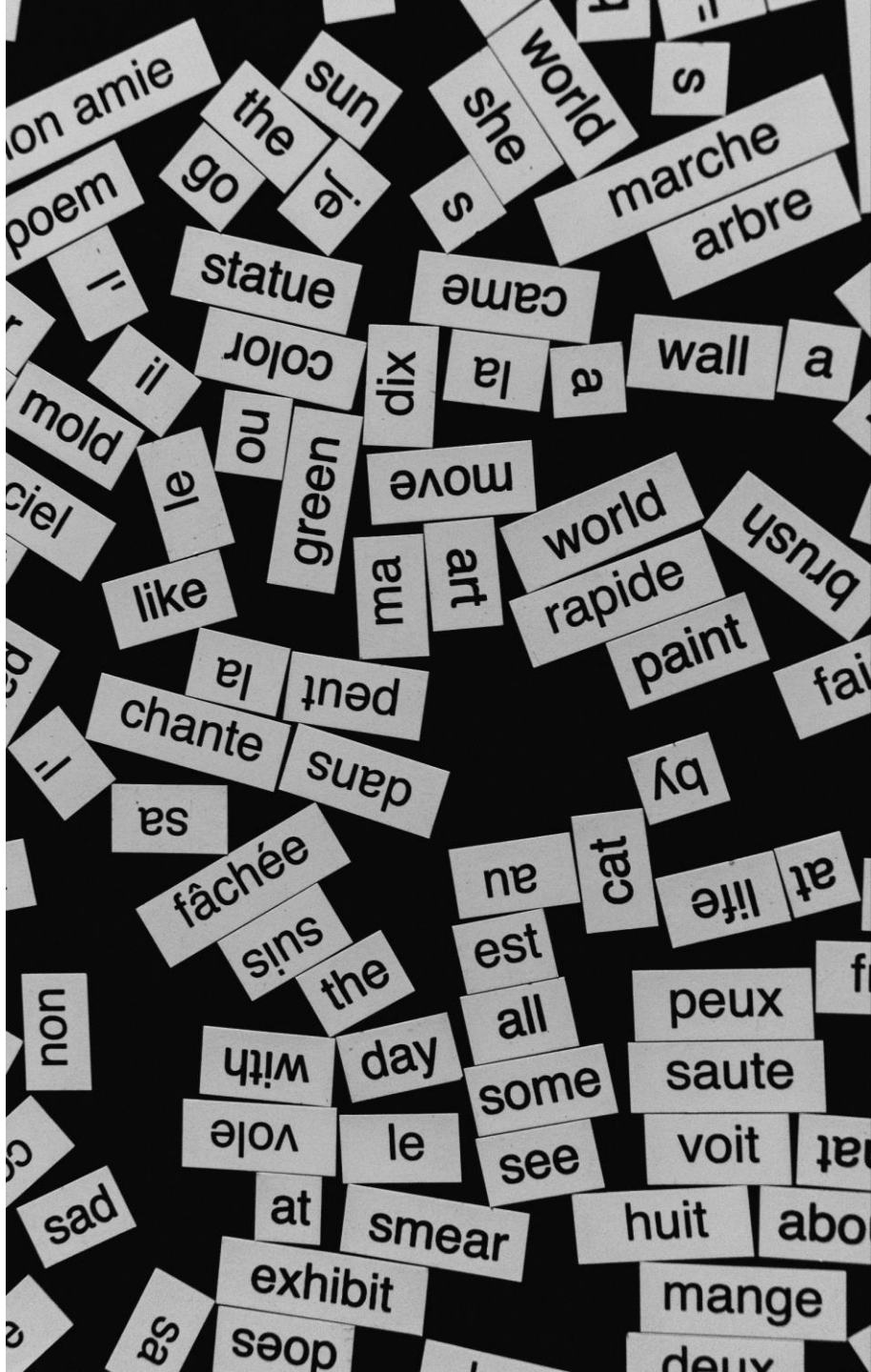
Data preprocessing

- ScienceDirect articles (2000–2025) on **emerging contaminants**
  - 13,705 articles in total
- Tokenization, lowercasing the tokens; removing punctuation, stop words, and numbers; stemming the tokens
  - vocabulary size of over 85,119 distinct words



Text processing

- LDA (Latent Dirichlet Allocation): to derive the terms that carry the main content of the text
  - roughly 40 different features (number of topics) in the entire corpus of all the journal abstracts
- Term Frequency–Inverse Document Frequency (TF–IDF): to compute vector representations for each abstract
  - a matrix of 13,705 rows (one row for each document) and 85,119 columns (one column for each distinct word)
- Singular Value Decomposition (SVD): to reduce dimensionality
  - The vector representations of the abstracts are high dimensional and sparse, thus introduce noise in the clustering
- Temporal (Time-Biased) clustering: to introduce temporal Bias
  - Time (year of publication) is added as a feature prior to clustering



Clustering

- K-means clustering (time-biased)
  - **Time Feature Inclusion**: Adds year of publication as a feature in document vectors.
  - **Bias Parameter**: Adjusts the weight of the time feature relative to other content features.
  - **Weight Adjustment**: Scales the importance of publication year to emphasize temporal grouping.
  - **Cluster Localization**: Forms clusters that are both thematically and temporally focused.
  - **Trend Detection**: Identifies topic emergence, peaks, and declines within specific periods.

Evaluation

- Trend Score Metric
  - **Silhouette Score:** to measure cluster cohesion; higher is better (range: -1 to 1)
  - **Standard Deviation of Years**: to assess temporal spread within clusters - lower SD means documents in a cluster are concentrated around specific time periods

$$\boldsymbol{Trend\ Score} = \frac{\boldsymbol{Silhouette\ score}}{\boldsymbol{SD_{years}}}$$

- **Purpose**: Identifies clusters that are both thematically coherent and temporally focused.
- **Outcome**: High Trend Score indicates effective trend detection in specific time frames.

Casualty Risk Consulting

Know You Can