applied sciences

MDPI

Article

Quantifying Fire Risk Index in Chemical Industry Using Statistical Modeling Procedure

Hyewon Jung [1], Seungil Ahn [1], Seungho Choi [1] and Yeseul Jeon [2,3,*]

[1] Insurance Data Team, Korean Fire Protection Association, Seoul 07328, Republic of Korea; hwjung@kfpa.or.kr (H.J.); asiih@kfpa.or.kr (S.A.); shchoi@kfpa.or.kr (S.C.)
[2] Department of Epidemiology & Biostatistics, University of California, San Francisco, CA 94143, USA
[3] Department of Statistics, Texas A&M University, College Station, TX 77843, USA
* Correspondence: jeons9677@tamu.edu

Abstract: Fire incident reports contain detailed textual narratives that capture causal factors often overlooked in structured records, while financial damage amounts provide measurable outcomes of these events. Integrating these two sources of information is essential for uncovering interpretable links between descriptive causes and their economic consequences. To this end, we develop a data-driven framework that constructs a composite Risk Index, enabling systematic quantification of how specific keywords relate to property damage amounts. This index facilitates both the identification of high-impact terms and the aggregation of risks across semantically related clusters, thereby offering a principled measure of fire-related financial risk. Using more than a decade of Korean fire investigation reports on the chemical industry classified as Special Buildings (2013–2024), we employ topic modeling and network-based embedding to estimate semantic similarities from interactions among words, and subsequently apply Lasso regression to quantify their associations with property damage amounts, thereby estimating the fire risk index. This approach enables us to assess fire risk not only at the level of individual terms, but also within their broader textual context, where highly interactive related words provide insights into collective patterns of hazard representation and their potential impact on expected losses. The analysis highlights several domains of risk, including hazardous chemical leakage, unsafe storage practices, equipment and facility malfunctions, and environmentally induced ignition. The results demonstrate that text-derived indices provide interpretable and practically relevant insights, bridging unstructured narratives with structured loss information and offering a basis for evidence-based fire risk assessment and management. The derived Risk Index provides practical reference data for both safety management and insurance underwriting by enabling the prioritization of preventive measures within industrial sites and offering quantitative guidance for assessing facility-specific risk levels in insurance decisions. An R implementation of the proposed framework is openly available for public use.

Keywords: fire incident analysis; risk index; text mining; chemical special buildings; network interaction

check for updates

Academic Editors: Wei Wang, Zhenxiang Tao and Tao Xu

Received: 26 September 2025
Revised: 24 October 2025
Accepted: 27 October 2025
Published: 28 October 2025

Citation: Jung, H.; Ahn, S.; Choi, S.; Jeon, Y. Quantifying Fire Risk Index in Chemical Industry Using Statistical Modeling Procedure. *Appl. Sci.* **2025**, *15*, 11508. https://doi.org/10.3390/app152111508

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fire accidents pose serious threats to industrial safety, human life, and the economy. In particular, chemical facilities are highly vulnerable because of their dependence on flammable and reactive materials, where minor ignition can escalate into catastrophic events. Understanding and quantifying such risks are therefore essential for developing

effective preventive measures and for guiding insurance and safety management policies. While structured fire statistics provide valuable information about fire frequency and damages, they often fail to capture the complex causal mechanisms recorded in textual fire investigation narratives. These reports, written by on-site officers, contain unstructured yet highly informative descriptions that reflect contextual factors such as process operations, equipment malfunctions, material leakage, and human or environmental interactions. Leveraging such unstructured narratives through text mining offers a promising pathway for discovering latent risk patterns that cannot be detected using aggregated statistics alone.

Previous research has employed a variety of approaches to analyze fire-related risks. Text-based studies have demonstrated the potential of mining unstructured documents. For instance, Kim and Hwang [1] applied topic modeling techniques to accident verdicts from ship fire cases to identify ignition sources, flammable materials, and negligence-related causes, while Tirunagari [2] investigated maritime accident investigation reports using text mining to extract causal relations among contributing factors. Although these studies represented important early steps in applying text analytics to fire accidents, they primarily relied on word-frequency clustering and did not capture the correlated structure among words that reflects the interdependent nature of fire risk factors. More recently, Liu et al. [3] analyzed the scientific literature to trace evolving research trends in forest fire studies, and Zhao et al. [4] constructed a comprehensive database of wildland–urban interface fires. Similar text mining frameworks have been used to identify risk factors in ship fires [5], heritage building fires [6], and industrial settings such as mining and chemical enterprises [7–11]. Narrative-based perspectives have also been explored: for example, Russo et al. [12] analyzed wildfire narratives to identify multiple social storylines concerning causes, consequences, and potential solutions, illustrating how unstructured narratives can provide a contextualized understanding of fire events and their socio-environmental implications.

Structured and indicator-based approaches have likewise received considerable attention. Ma et al. [13] conducted a data-driven analysis of over one million building fire reports, integrating structured incident records with socioeconomic and structural variables to assess the effects of detection systems and extinguishing devices on fire spread and injury risk. Similarly, Zhang et al. [14] developed an indicator system for urban fire risk assessment that emphasizes meteorological and building characteristics. Other domain-specific efforts include quantitative frameworks for Value-at-Risk of heritage structures [6], climate-related hazard assessment [15], and chemical industry safety evaluation [16]. However, most of these structured approaches exclude unstructured textual data, thereby omitting rich qualitative details that often contain the causal reasoning behind fire events.

Parallel progress in natural language processing and machine learning has broadened the methodological basis for text-based risk analytics. Transformer-based models such as Sentence-BERT have improved short-text clustering by modeling contextual similarity between sentences [17,18], while hybrid topic–autoencoder architectures enhance the detection of latent thematic structures [19,20]. Graph-based and network-analytic frameworks have also been proposed to represent relationships among semantic entities, enabling the discovery of interaction mechanisms in multilingual or domain-specific corpora [21–23]. These methodological advances have improved text clustering, anomaly detection, and classification across diverse fields, including risk disclosure analysis [24], and insurance claim interpretation [25]. However, existing fire-related applications remain largely descriptive, focusing on frequent keywords or co-occurrence patterns rather than connecting textual semantics to measurable outcomes such as financial or property loss.

Taken together, the literature reveals a persistent methodological and practical gap. Existing fire text-mining studies seldom model the dependent structure among risk-related

terms that jointly describe operational and environmental conditions within accident narratives. Likewise, few studies link textual features to structured indicators such as damage costs or insurance data. Consequently, current approaches cannot quantify how textual expressions of risk translate into tangible losses or prioritize critical risk factors by their financial relevance. Addressing this gap, the present study introduces a unified framework that systematically analyzes fire investigation narratives, estimates semantic dependencies among risk-related terms, and integrates these text-derived features with estimated damages data to construct a composite, loss-aware fire Risk Index. This approach bridges unstructured and structured domains, offering a scalable and interpretable foundation for quantitative fire risk analysis and decision-making.

This framework is expected to be integrated into industrial safety management and insurance risk-assessment systems. The derived Risk Index provides quantitative reference data that enable facility managers to prioritize preventive measures according to site-specific hazards extracted from textual records. It also assists insurers in evaluating facility-level risks that cannot be captured by conventional loss-ratio statistics. By linking narrative investigation data with measurable financial outcomes, the framework facilitates data-driven decision making for both workplace safety management and insurance underwriting.

Contributions

This study makes the following contributions:

- First large-scale analysis of Korean fire investigation narratives. To the best of our knowledge, this is the first systematic attempt to analyze over a decade (2013–2024) of textual records on the chemical industry, classified as Special Buildings, written by fire officers. These narratives provide direct accounts of ignition sources and causal conditions that have not been previously examined in quantitative fire safety research.
- Integration of textual evidence with economic loss indicators. By linking unstructured narratives with structured data on fire property damage, we move beyond frequency-based measures of risk. This integration enables the identification of risk factors that matter not only for their occurrence, but also for their economic impact, thereby offering a more practical and policy-relevant assessment of fire risk.
- Development of a statistically grounded risk index. We combine established statistical approaches [26] with a Lasso regression framework to construct a novel composite index. This allows us to capture meaningful dependency structures among words and quantify their contribution to observed loss outcomes. The resulting index provides an interpretable and efficient tool to quickly identify critical fire-related terms that elevate risk.

The remainder of this paper is organized as follows: Section 2 introduces the dataset and outlines the analytical procedures, including latent topic estimation via the Biterm Topic Model, topic clustering through a Latent Space Item Response Model, and the construction of a Risk Index factor using Lasso regression. Section 7 presents the results, covering topic characterization through words, thematic aggregation of topics, and evaluation of the proposed Risk Index factor. Section 8 discusses the implications of the findings and concludes the study.

2. Materials and Methods

Figure 1 presents the proposed analytical framework for constructing a fire risk index from unstructured fire investigation texts. In the first step (Step 1), natural language processing techniques are applied to preprocess the textual records and extract nouns, with a particular emphasis on building a domain-specific lexicon related to the chemical

industry. This enables the identification and expansion of keywords that are directly relevant to chemical processes and accident scenarios, providing a structured corpus for further analysis.

In the second step (Step 2), a Biterm Topic Model [27] is employed to classify documents into latent topics and to estimate the distribution of words within each topic. This step not only offers a compact summary of large-scale documents, but also transforms unstructured text into a topic–words distribution matrix that can be further exploited. Using these topic–word distributions as input for the subsequent latent space model is advantageous, as it embeds words into a representation that reflects their co-occurrence patterns, thereby facilitating the estimation of meaningful word–word interactions that are not apparent from raw text alone.

In the third step (Step 3), the latent item response model [28] is applied to infer the positions of words in a continuous latent space, which captures their semantic relationships. The estimated distances between words can be interpreted as measures of semantic similarity: for example, if two words are placed close to each other in the latent space, they are more likely to represent semantically related concepts. Leveraging these latent positions, words are clustered into semantically coherent groups, thereby enabling the construction of interpretable clusters of risk-related vocabulary. Importantly, the estimation of interword interactions accounts for both the frequency and contextual sparsity of co-occurrences. Thus, words are not measured as strongly associated merely because they occasionally appear together; rather, their latent positions are inferred based on consistent and dense interaction patterns with other words across the corpus.

In the final step (Step 4), structured data on property damage amounts are incorporated into the analysis. Specifically, Lasso regression is used to estimate coefficients linking each word to the magnitude of property loss. Words with higher coefficients can thus be interpreted as risk factors associated with greater expected damages. Beyond word-level inference, this integration allows for cluster-level analysis: by examining the aggregated coefficients of words within each cluster, we can identify which semantic groupings correspond to high-risk factors in terms of potential financial losses. Taken together, this framework not only provides a systematic way to quantify fire risk from unstructured narratives, but also bridges semantic information extracted from text with structured loss data to yield interpretable and practically meaningful risk indices.

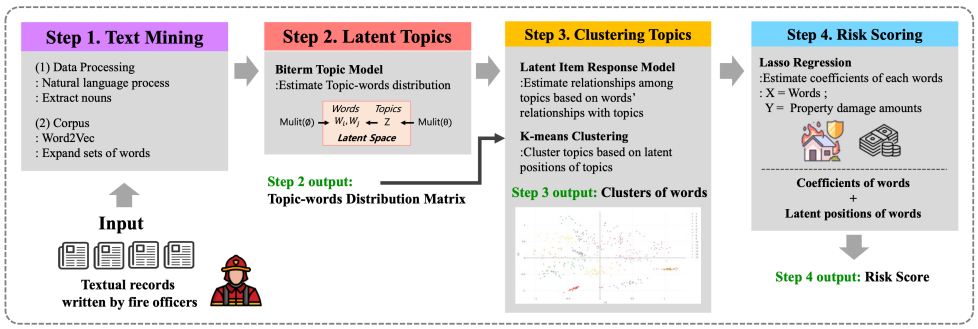


Figure 1. Analytical framework for constructing a fire risk index from unstructured investigation texts. The process begins with text mining to preprocess investigation records and extract key nouns (Step 1). A Biterm Topic Model is then applied to estimate topic–word distributions (Step 2). Next, a latent item response model is used to infer topic–word interactions and to estimate the latent positions of words, which are subsequently clustered to capture semantic groupings (Step 3). Finally, Lasso regression estimates the coefficients of words using property loss data as the outcome, yielding a composite Risk Index that integrates words' semantic closeness with structured information derived from loss amounts (Step 4).

3. Data

This analysis focuses on fire accidents in Special Buildings in Korea that fall under the chemical industry (excluding plastic production) within the category of manufacturing uses. In Korea, Special Buildings are legally designated facilities identified as high fire risk based on their intended use, scale, and other criteria. These facilities are overseen by the Korea Fire Protection Association (KFPA) for fire prevention purposes. As of the end of 2024, there were 54,517 such buildings nationwide, and the total number changes annually.

The dataset used in this study consists of records from post-incident investigations of fire accidents in these chemical industry Special Buildings between 2013 and 2024. As shown in Table 1, the fire investigation records are unstructured text documents describing the fire circumstances, such as the cause, involved equipment, and ignited materials. All investigation texts are written in Korean. In addition, the dataset includes property damage estimates verified by local fire departments in accordance with the *Fire Damage Assessment Manual of the National Fire Agency of Korea*, covering both movable and immovable property losses such as buildings, facilities, machinery, and inventory assets.

Table 1. Ten randomly selected representative excerpts of fire investigation records.

ID	Fire Investigation Records	Damage Estimates (USD)
1	A fire occurred in the vulcanization process of Plant 2. Flames and smoke were first observed under the pit of equipment No. 3115 on the 3100 line. The area showed severe damage and electrical abnormalities, and CCTV confirmed early smoke from the same location. The fire is presumed to have originated beneath the pit of No. 3115.	$126,676,000
2	In the production building of a factory, sunflower oil was being refined by mixing oil, hexane, and activated carbon with a stirrer. Hexane vapors leaked and ignited from an unidentified ignition source, causing an explosion with fire.	$4,471,000
3	During repair on the duct connected to a dust collector, grinder sparks ignited sponge soundproofing material on the wall. Scene investigation indicated initial ignition at the wall near the duct; burn patterns showed vertical spread to the second floor. The cause was determined as worker negligence.	$3,566,000
4	Prior to an explosion and fire, an emergency call reported injuries due to steam leakage. Investigation found that unreacted photoinitiator was normally returned to a 5 ton reactor, but aluminum chloride catalyst was mistakenly added through a manhole, causing ejection. While workers evacuated, solvent vapors accumulated and later ignited from an unknown source, causing an explosion and fire.	$3,547,000
5	A witness heard a beep and saw vapor leaking from a reactor lid. After donning a respirator, larger vapor was seen from the lid, condenser, and receiver tank area; droplets and chemicals fell from the reactor bottom. A brief initial explosion occurred, followed by a major explosion seconds later.	$3,053,000
6	In a pharmaceutical production plant, an explosion and fire occurred during blending and drying of raw materials in the third floor mixing and granulation room. The fire spread and damaged a large portion of equipment and facilities.	$1,808,000
7	A three story chemical plant producing BP F 95 experienced a fire presumed to have originated inside a centrifugal dehydrator while dehydrating BP F 95 mixed with toluene. The fire partially spread up to the third floor; no injuries were reported.	$1,298,000
8	An explosion occurred due to a reaction between reactor contents and unidentified static electricity in a synthesis plant. The fire spread to nearby organic solvents, resulting in total destruction of the synthesis building.	$1,287,000
9	In the drying room of Building D, an operator was loading wet powder cosmetic raw material into a dryer tray. Acetone vapors reached an explosive range and presumably ignited within the operator's working radius, spreading to nearby combustibles.	$967,000
10	No other heat source was identified at the origin except sparks from welding work. Burn patterns indicated rapid spread from stored flammable material (styrofoam). Based on witness statements and proximity to the welding site, welding sparks were presumed to have ignited the styrofoam.	$942,000

By leveraging more than a decade of investigation reports (2013–2024), this analysis captures long-term and stable patterns of fire incidents within a relatively homogeneous industrial environment. This extensive temporal coverage mitigates sample-specific variability and enhances the generalizability of the derived semantic structures.

A total of 8190 fire investigation reports were analyzed, from which 449 representative keywords were selected based on three filtering criteria: excluding 763 non-informative terms (e.g., numbers, symbols, general words, or names), consolidating 315 synonymous words into 165 representative entries, and retaining only those appearing at least twice across all reports.

4. Latent Topic Estimation via the Biterm Topic Model

We employed the Biterm Topic Model (BTM) to uncover latent semantic structures within textual statements written by firefighters during post-incident investigations of fire causes. As a preprocessing step, morphological analysis was conducted to decompose words into their base morphemes. From these, we extracted nouns in their canonical form, which constituted the initial corpus. To enrich this corpus, we expanded the vocabulary set using the `word2vec` algorithm, which maps words into a Euclidean latent space according to semantic similarity. By identifying words in close proximity within this embedding space, we obtained an augmented corpus that better captured the semantic landscape of the texts.

The ultimate objective was to identify a collection of latent topics that summarize the semantic content of the documents. Each topic was represented as a distribution over words, while each document was modeled as a mixture of these latent topics. Since those reports typically consist of fewer than 200 words and can thus be regarded as short texts, the BTM was particularly suitable for this setting. The model relied on extracting biterms, i.e., unordered word pairs within documents, which served as the input to the topic model.

The BTM is founded on several key assumptions:

- Each pair of words (biterm) is assumed to arise from an underlying latent topic.
- Topics themselves represent semantically coherent clusters of words.
- Word co-occurrence patterns within the corpus can therefore be explained through mixtures of such latent topics.

Formally, the likelihood of BTM is determined by the topic distribution and the topic–word distributions. Two sets of parameters must, therefore, be estimated: the topic proportion vector $\boldsymbol{\theta}$ and the topic–word distributions $\boldsymbol{\phi}_z$. The prior for each $\boldsymbol{\phi}_z$ is specified as a Dirichlet distribution with hyperparameter β, while the prior for $\boldsymbol{\theta}$ follows a Dirichlet distribution with hyperparameter α. A latent topic assignment variable z is drawn from a Multinomial distribution with parameter $\boldsymbol{\theta}$, and, conditional on z, each word is generated from $\boldsymbol{\phi}_z$. Hence, the parameters of interest are $\boldsymbol{\theta}$, $\boldsymbol{\phi}_z$, and z.

The generative process of BTM can be summarized as follows:

Step 1 Draw a topic distribution $\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha)$.
Step 2 For each biterm $b \in B$, assign a latent topic $z \sim \text{Multinomial}(\boldsymbol{\theta})$.
Step 3 For each topic z, draw a topic–word distribution $\boldsymbol{\phi}_z \sim \text{Dirichlet}(\beta)$.
Step 4 Generate the two words $w_i, w_j \sim \text{Multinomial}(\boldsymbol{\phi}_z)$.

The Biterm Topic Model (BTM) estimates the probability of observing word pairs (biterms) across the entire corpus, rather than within individual documents. The resulting joint likelihood over all biterms B is expressed as

$$p(B) = \prod_{i,j} \sum_{z} \boldsymbol{\theta}_z \, \boldsymbol{\phi}_{i|z} \, \boldsymbol{\phi}_{j|z}. \tag{1}$$

This formulation allows for the model to capture global word co-occurrence patterns that help infer more stable topic–word associations. The conditional posterior for assigning a biterm to topic assignment z is given by

$$p(z|\mathbf{z}_{-b}, \mathbf{B}, \alpha, \beta) \propto (n_z + \alpha)\frac{(n_{w_i|z} + \beta)(n_{w_j|z} + \beta)}{(\sum_w n_{w|z} + M\beta)^2}, \quad (2)$$

where n_z is the number of biterms currently assigned to topic z, $n_{w|z}$ is the number of times word w is assigned to topic z, and $\mathbf{z}_{-b}$ denotes topic assignments excluding the current biterm. Intuitively, this expression increases the probability of assigning a biterm to topics that already have high counts for its component words, while α and β serve as smoothing hyperparameters controlling topic and word diversity.

To address convergence issues that may arise from direct Gibbs sampling, we employ collapsed Gibbs sampling, which integrates $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_z$ by exploiting the conjugacy of the Dirichlet–Multinomial distributions [29]. The algorithm of the Biterm Topic Model (BTM) is presented in Algorithm 1.

After sufficient sampling iterations, the posterior estimates of topic–word distributions and topic proportions are

$$\boldsymbol{\phi}_{w|z} = \frac{n_{w|z} + \beta}{\sum_w n_{w|z} + M\beta}, \quad \boldsymbol{\theta}_z = \frac{n_z + \alpha}{|B| + K\alpha}. \quad (3)$$

Algorithm 1 Collapsed Gibbs Sampler for BTM

Input: number of topics K, hyperparameters α, β, biterm set $\mathbf{B}$
Output: topic–word distributions $\boldsymbol{\phi}_{w|z}$ and topic proportions $\boldsymbol{\theta}_z$

```
Initialize topic assignments randomly for all biterms
for iteration = 1, 2, ..., N do
    for each biterm b ∈ B do
        Sample z_b from p(z|z_{-b}, B, α, β)
        Update counts n_{w|z} and n_z
        Compute parameters φ_{w|z} and θ_z
    end for
end for
```

Here, M denotes the total number of unique words in the corpus and K represents the number of latent topics. Accordingly, the terms $M\beta$ and $K\alpha$ act as normalization constants arising from the Dirichlet–Multinomial conjugacy, where β and α control the concentration of word and topic distributions, respectively. Larger values of β or α lead to smoother, less peaked distributions, while smaller values encourage more concentrated topic–word associations.

Specifically, we constructed a word–topic probability matrix $\mathbf{X}$ of dimension $N \times P$, where N is the number of words and P is the number of topics. To identify characteristic words, we computed two measures for each row of $\mathbf{X}$: (i) the coefficient of variation across topics, and (ii) the maximum probability. The coefficient of variation captures the relative dispersion of a word's probabilities across topics, while the maximum probability identifies whether the word is strongly associated with at least one topic. Words with both high dispersion and high maximum probability are retained as representative terms, enabling more interpretable characterization of latent topics.

5. Clustering Topics via Latent Space Item Response Model

We estimate interactions among topics and visualize their relationships by embedding them into a latent interaction map. To achieve this, we follow the approach of Jeon et al. [26], who applied LSIRM to topic–word distributions, and employ its Gaussian version in our setting. Specifically, we use the Gaussian LSIRM to represent the bipartite structure

between topics and words, where topics are regarded as "items" and words as "respondents" as below:

$$x_{i,p} \mid \Theta = a_i + b_p - ||\mathbf{v}_i - \mathbf{u}_p|| + \epsilon_{i,p}, \quad \epsilon_{i,p} \sim \mathrm{N}(0, \sigma^2), \tag{4}$$

where $x_{i,p}$ denotes the probability of word i belonging to topic p. The parameters a_i and b_p represent the intrinsic activity of word i (i.e., how generally a word tends to appear across topics) and the popularity or overall strength of topic p, respectively. The Euclidean distance $||\mathbf{v}_i - \mathbf{u}_p||$ measures how semantically close the word and topic are in the latent space, where smaller distances indicate stronger associations. Bayesian inference is employed to estimate the full parameter set $\Theta = \{a, b, \mathbf{U}, \mathbf{V}\}$ with appropriate priors, and parameters are sampled using Markov chain Monte Carlo (MCMC). As a result, we obtain latent positions of topics $\mathbf{v}_i$ in $\mathbb{R}^d$, forming the topic coordinate matrix $\mathbf{A} \in \mathbb{R}^{d \times P}$.

Once the latent positions of the topics were estimated, we proceeded to cluster the topics in order to identify groups with similar semantic characteristics. To this end, we applied the K-means clustering algorithm to the latent position matrix $\mathbf{A}$. The K-means method partitions the P topics into C disjoint clusters $\{C_1, \ldots, C_C\}$ by minimizing the within-cluster sum of squared distances:

$$\arg\min_{C_1,\ldots,C_C} \sum_{c=1}^{C} \sum_{\mathbf{v}_i \in C_c} \|\mathbf{v}_i - \mu_c\|^2, \tag{5}$$

where μ_c denotes the centroid of cluster C_c. This clustering step groups topics that are located close together in the latent space, thereby reflecting their semantic similarity, as derived from the topic–word distributions.

The resulting framework not only enables us to visualize the relationships among topics through their latent embeddings, but also to categorize them into interpretable clusters. Specifically, LSIRM provides a probabilistic mechanism to embed topics in a common latent space, and K-means clustering on the estimated positions further organizes these topics into coherent groups. This joint approach allows us to explore both the global structure (via the interaction map) and the local grouping (via cluster assignments) of topics inferred from the text data.

6. Risk Index Factor via Lasso Regression

To assess the contribution of words in explaining fire-related financial losses, we modeled the expected property damage amount y for each incident report as the outcome variable and the extracted words as predictors x. Specifically, let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the vector of estimated damages across n reports, and let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the document–term matrix, where each column corresponds to one of the p words. We fit a Lasso regression model [30] of the form

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{6}$$

where $\lambda > 0$ is a tuning parameter. The ℓ_1 penalty induces sparsity in the estimated coefficients, effectively performing variable selection by shrinking many coefficients toward zero. This property enables us to identify only those words that exhibit substantial predictive power for property damage amounts.

The estimated coefficients $\hat{\beta}_j$ quantify the influence of word j on expected fire-related losses. Coefficients close to zero indicate little or no association between the word and the damage amount, while coefficients with larger magnitudes capture stronger effects. Moreover, the sign of $\hat{\beta}_j$ allows for direct interpretation: negative values imply that the

presence of a word is associated with lower property damage, whereas positive values suggest that the word signals higher expected losses. In this way, the Lasso framework provides a principled mechanism to evaluate and interpret the importance of words in the context of fire incident reports. Building on the estimated coefficients, we further construct a Risk Index to quantify word-level contributions in a structured manner. Specifically, we derive the index along three complementary dimensions.

First, within each cluster of semantically related words, the estimated coefficients are rescaled using a min–max transformation to lie between 0 and 1. This yields the *(i) Risk Index within Cluster* ($\gamma_{i,c}, i = 1, \cdots, n_c, c = 1, \cdots, C$), which allows for comparison of words relative to others in the same group. Second, to assess risk at the cluster level, we compute the mean coefficient value across words in each cluster c and again apply min–max scaling to obtain a *(ii) Cluster-Level Risk Index* between 0 and 1 ($\delta_c, c = 1, \cdots, C$). This provides an interpretable measure of the overall risk associated with each cluster. Finally, the *(iii) Overall Risk Index* ($\boldsymbol{\rho}_{i,c}$) for a word is defined as the average of its within-cluster score and the risk score of its corresponding cluster. In this way, the framework jointly accounts for both the local importance of a word relative to its peers and the broader risk tendency of its semantic group. The resulting index serves as a principled indicator to identify the key linguistic factors associated with higher fire-related property damages in incident reports.

7. Results

7.1. Topic Characterization Through Words

As illustrated in Step 2 of Figure 1, the fire investigation texts were first analyzed using topic modeling to identify latent semantic structures. Table 2 presents the resulting thematic groups, providing an overview of how the documents are organized into distinct topics. In addition, Table 3 presents the distribution of words derived from the estimated topic–word probabilities ($\boldsymbol{\phi}_{w|z}$) obtained in the topic modeling stage. The table lists the top five terms with the highest posterior probabilities for each topic, highlighting representative keywords that characterize the semantic content of each domain.

Table 2. Mapping of extracted topics from fire investigation records to their descriptive names. The thematic categories are assigned based on high-probability keywords (summarized in Table 3) and domain-specific interpretation.

Topic	Name
Topic 1	Flammable vapor ignition due to the use of organic solvents such as cleaning solutions
Topic 2	Explosions caused by sparks generated from friction or static electricity due to the blending of flammable or combustible raw materials
Topic 3	Fires caused by common electrical factors
Topic 4	Ignition caused by sludge accumulation within ventilation equipment
Topic 5	Ignition caused by electrical heating
Topic 6	Related to dust collection equipment
Topic 7	Spontaneous combustion caused by improper storage or containment of flammable residues
Topic 8	Fire caused by forklift operation
Topic 9	Chemical explosion occurring in the reactor
Topic 10	Ignition of residue accumulated in ducts connected to or adjacent to dust collection equipment
Topic 11	Due to improper use of drying equipment
Topic 12	Fire caused by improper use of a banbury mixer for rubber molding
Topic 13	Combustible materials ignited due to welding spark during hot work
Topic 14	Fire caused by ignition sources in machinery with operation motor, such as air compressors
Topic 15	Ignition of waste materials due to improper disposal of cigarette butts

Topic 1 reflects the incidents associated with oil vapors generated from organic solvents or related chemical processes, frequently occurring in cleaning operations or wastewater treatment facilities. Topic 2 captures the fire risks arising from the ignition of combustible materials due to friction or static electricity during equipment operation, as suggested by keywords such as *mixer* and *drum can*. Topic 3 represents general electrical fires, character-

ized by terms such as *electrical short circuit* and *circuit breaker*, while Topic 5 is more narrowly related to electrical heating sources (e.g., *air conditioners* and *heating wires*). Topic 4 emphasizes ignition triggered by sludge accumulation within ventilation systems, particularly in laboratory environments.

Topic 6 highlights the fires directly linked to dust collection equipment, where flames may propagate through ducts or filter systems containing combustible dust. In contrast, Topic 10, although conceptually related, places less emphasis on dust collection equipment itself, and instead indicates fire hazards involving adjacent facilities, such as *ventilation ducts*, *plastics*, and *drying equipment*. Topic 7 is dominated by spontaneous combustion events caused by the improper storage of combustible residues, including processed byproducts such as *sesame dregs*. Topic 8 involves forklift-related fires, which occur across both chemical and general factory settings, often due to battery or engine compartment failures. Topic 9 concerns reactor-related incidents in chemical plants, where abnormal reactions generate oil vapors leading to explosions.

Other topics describe equipment-specific or context-specific fire causes. Topic 11 focuses on drying equipment, particularly in cases involving powders such as *silicon*. Topic 12 reflects fires ignited by the thermal oil of *banbury* mixing equipment. Topics 13–15 capture more general factory-related accidents: Topic 13 refers to welding-induced ignition near cooling towers or sandwich panels; Topic 14 highlights motor-related electrical fires in air compressors; and Topic 15 illustrates landfill or waste-area fires, frequently initiated by discarded cigarette butts.

While biterm modeling provides an effective means of partitioning documents into topics, it assigns every word to all topics, which limits its ability to capture direct relationships among words. Since the primary goal of this study is to identify the words that carry substantial meaning within fire investigation documents, it is essential to explore how words are grouped through their interactions. In this respect, topic–word distributions offer probabilistic information on the degree of association between words and topics, thereby serving as a valuable resource for indirectly inferring inter-word relationships. Leveraging this information enables us to examine documents not only at the level of topics, but also from the perspective of word-level associations, ultimately facilitating a more interpretable summarization of fire-related narratives.

Table 3. Top five representative words for each extracted topic based on topic word probability distributions. These high probability words provide the basis for interpreting the thematic characteristics of the topics.

Topic	Top 5 Words
Topic 1	cleaning, cleaning room, wastewater treatment plant, agitator, machine room
Topic 2	mixer, drum can, pallet, plastic, explosion
Topic 3	electrical short circuit, electrical distribution, flame, circuit breaker, SWGR
Topic 4	extractor hood, oven, laboratory, transformer, electricity
Topic 5	heating wire, outdoor unit of air conditioner, thermal/acoustical insulation, air conditioner, rooftop
Topic 6	dust collection equipment, duct, flame, plenty of, filter system
Topic 7	storage, sesame dregs, residues, spontaneous combustion, storage room
Topic 8	forklift, battery, electrical short circuit, distribution, engine compartment
Topic 9	explosion, container, reactor, oil vapor, static electricity
Topic 10	dust collection equipment, duct, flame, ventilation duct, plastic
Topic 11	drying equipment, powder, silicon, base material, storage
Topic 12	base material, mixing equipment, Banbury, thermal oil, flame
Topic 13	welding, cooling tower, flame, sandwich panel, acetylene
Topic 14	air compressor, motor, electrical short circuit, vulcanizer, solenoid valve
Topic 15	waste materials, cigarette butts, waste, stacking, flame

7.2. Thematic Aggregation of Topics

Building upon the topic–word distributions, we further inferred interactions among words to explore higher-level thematic structures. Figure 2 presents a two-dimensional projection of the latent positions of the 449 words introduced in Section 3, with clusters identified using the *k*-means algorithm. The visualization highlights clusters by distinct colors, showing how semantically related words are grouped in close proximity. Based on model selection criteria, a total of 15 clusters were identified as valid. As shown in Figure 2, words located near one another in the latent space tend to form coherent clusters, thereby capturing meaningful associations beyond the topic-level representation.

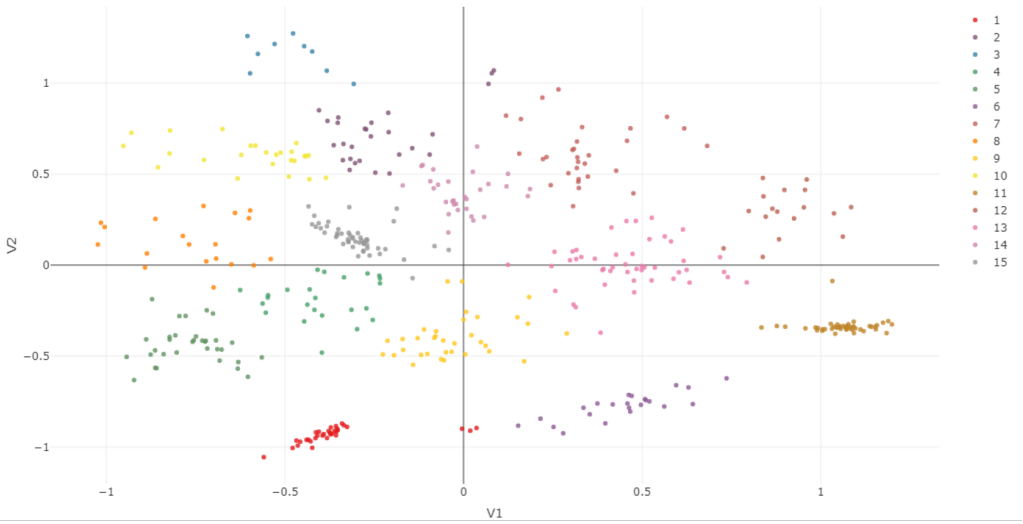


Figure 2. Two-dimensional latent representation of words with clustering results obtained using the *k*-means algorithm. Distinct colors indicate the 15 identified clusters, showing how semantically related words are grouped in close proximity.

The interpretation of clusters requires careful consideration because the meaning of a single word is better understood in relation to its neighboring terms. A word located at the center of a cluster typically exhibits a high probability of contributing to the generation of a topic, often alongside other words in the same group. However, the semantic similarity within a cluster may arise either from the joint contribution of the words themselves or from their shared functional context with alternative terms. Thus, rather than examining individual words in isolation, it is crucial to evaluate their surrounding vocabulary to delineate the collective meaning of each cluster. In this sense, clusters serve as categorical units in which semantic coherence emerges from localized word proximities, allowing for the clarification of latent thematic structures.

Following Step 3 in Figure 1, the estimated latent positions of words were used to cluster semantically related terms based on their pairwise distances. Table 4 presents the top 10 representative keywords for each cluster, arranged according to their proximity to other words within the same cluster. This ordering enables a more precise characterization of the distinctive features of each cluster. While the identified clusters are not directly labeled as specific fire causes, their semantic coherence often reflects shared causal contexts. Accordingly, these clusters can be interpreted as data-driven representations of latent cause structures underlying fire incidents. Cluster 1 is anchored by the central term *distillation column*, which suggests the risk of explosion due to chemical leaks or failures in pressure control. Surrounding words such as *exposure*, *solvent*, *heptane*, and *Silanes manufacturing* reinforce the theme of hazardous material leakage leading to fires or explosions. Accordingly, Cluster 1 is labeled as *Fire or explosion caused by hazardous material leakage*.

Cluster 2 excludes the generic fire-related term *ignition source* and instead centers on *polishing*, which can produce ignition hazards when frictional heat contacts combustible materials. Neighboring words such as *painting*, *floor*, *waste*, and *interior wall* indicate that this cluster corresponds to *Ignition from heat accumulation near combustible interior materials*.

Cluster 3 is defined by the term *incinerate*, reflecting the inherent fire risks of incineration processes. Associated words such as *waste wood*, *interior materials*, *heat*, and *urethane* highlight ignition due to residual heat, while *cutting machines* and *dust collection equipment* suggest mechanical sources of smoldering. Thus, Cluster 3 is interpreted as *Ignition due to residual heat post-work with combustibles*.

Cluster 4 revolves around the term *heat of reaction*, but further interpretation requires contextualization with neighboring terms such as *expired reagents*, *corn*, and *cooking oil*. These materials, when awaiting disposal, pose risks of spontaneous combustion, particularly when combined with *oxygen*, *heat waves*, or *rainwater*. Hence, this cluster represents *Spontaneous combustion from abandoned chemicals and oils*.

Cluster 5 is structured around *abnormal reaction*, which links to numerous other terms and signals fires or explosions during abnormal chemical processing. Words such as *pharmaceutical*, *film*, *oil*, and *epoxy* point toward industrial settings where chemical instability can result in severe accidents. This cluster is labeled as *Fires or explosions from abnormal reactions during chemical handling*.

Cluster 6 includes terms like *storage*, *cable*, *control box*, and *electrical circuit board*, reflecting general electrical fires not tied to specific chemical processes. Its theme is summarized as *Electrical fires in areas handling flammable materials*.

Cluster 7 is characterized by *coating*, which signals risks associated with flammable paints. Surrounding terms such as *corrosion*, *cutting oil*, *grinders*, and *presses* suggest ignition by sparks or friction from industrial machinery. Accordingly, Cluster 7 is described as *Ignition of flammable substances (e.g., machine oil) by friction heat or spark from machinery*.

Cluster 8 is centered on *vapor*, a strong indicator of fire hazards in chemical plants. Terms such as *nucleic acid*, *grease*, and *hazardous materials* highlight risks arising from the ignition of volatile organic vapors, supporting the interpretation of *Ignition of flammable vapor during organic solvent use*.

Cluster 9 highlights the term *drying room*, referring to environments with elevated fire risks due to sustained high temperatures. Neighboring words, including *cosmetic*, *ventilation fans*, *rotating*, *small amount*, and *decompose*, suggest fire hazards associated with oil vapor and production processes. Thus, Cluster 9 is classified as *Ignition from temperature rise in oil vapor areas*.

Cluster 13 incorporates *dilution*, *power outage*, *tracking*, *muller*, *vulcanizer*, and *melting fusion*, suggesting interactions between hazardous materials, electrical ignition sources, and mechanical heat. This combination indicates *Ignition in areas with accumulated combustible dust and oil vapors*.

Cluster 14 includes terms such as *micro*, *seat pad*, *deodorizing tower*, *absorbent pad*, and *gunnysack*, which are not individually definitive. However, the presence of *compost* and *sesame dregs* highlights substances prone to spontaneous combustion through oxidation heat. Accordingly, this cluster represents *Spontaneous combustion from improper oil residues storage or disposal*.

Finally, Clusters 10, 11, 12, and 15 are each characterized by distinct keywords: Cluster 10 corresponds to *Fire from ignition in accumulated combustibles*, Cluster 11 to *Fires from electrical factors in process equipment sites*, Cluster 12 to *Fires during machinery maintenance*, and Cluster 15 to *Ignition from sparks inside dust collection equipment with filters and debris*. Importantly, these categories represent fire scenarios that extend beyond chemical plants to general factory environments.

Table 4. Clusters of ten representative keywords derived from inter-word correlations based on estimated latent positions.

Cluster	10 Words
1	distillation column, exposure, solvent, photoinitiator, metal, cleaning solvent, heptane, nozzle, flexible, silanes manufacturing
2	ignition source, polishing, painting, floor, laboratory, electricity, plenty of, scrap paper, waste, recycling
3	incinerate, waste wood, interior material, heat, urethane, smoldering ignition, cutting machine, dust collection equipment, rubber, manufacturing machine
4	heat of reaction, expired reagent, corn, cooking oil, storage tank, reagent, oxygen, eruption, heat wave, rainwater
5	abnormal reaction, pharmaceutical, film, sheath heater, upper, suction, oil, wire mesh, epoxy, dust explosion
6	storage, cable, malfunction, cool down, oil tank, control box, coil, electrical circuit board, refrigerator, fertilizer
7	coating, injection, corrosion, cutting oil, grinder, press, repair, acetylene, scrubber, steam equipment
8	vapor, nucleic acid, grease, leak, hazardous material, mix, paint, high temperature, large scale, thermal cutting
9	drying room, cosmetic, pallet, ventilation fan, pressure, ceiling, rotation, small amount, decompose, lid
10	preheat, ignition, stacked, accumulation, tire, moisture, activated carbon, semi-finished product, waste storage, welding
11	cooling fan, service line, shut out, lower terminal, insulating oil, underground, fan belt, switch, electric current, analysis lab
12	closing, charging equipment, flammability, maintenance, automation, repair, cooling machine, machine room, gas torch, arc
13	dilution, power outage, cover, immediate upper, tracking, prefabricated, vacuum, muller, vulcanizer, melting fusion
14	micro, seat pad, deodorizing tower, absorbent pad, gunnysack, ton bag, compost, boiler room, sesame dregs, oxidation heat
15	belt, particle, capture, aluminum, inlet, rubber department, wood chip, blowing, metal powder, abradant

Taken together, the interpretation of clusters provides meaningful categorical insights into the semantic structure of fire investigation records. However, semantic similarity alone does not capture the extent of economic severity associated with each term. To address this, we incorporated fire damage estimates into the analysis by estimating word-level coefficients via LASSO regression. This approach allows us to link clusters with the magnitude of potential financial losses. The coefficients estimated for each word quantify its relative contribution to explaining variations in property damage estimates. Figure 3 presents a three-dimensional visualization, where the horizontal axes represent the latent positions of words and the vertical axis corresponds to their regression coefficients. Each colored dot denotes a single word, and the color indicates its cluster membership consistent with the grouping shown in Figure 2. This visualization allows for direct comparison between semantic proximity (latent positions) and the estimated contribution of each word to financial damage. Words located at higher positions along the coefficient axis correspond to terms associated with higher risk of financial loss, whereas clusters with generally lower coefficient values indicate groups of words linked to less severe incidents. By examining the clusters from multiple viewing angles, one can observe that several clusters exhibit a wide range of coefficient magnitudes, suggesting heterogeneous risk levels within the same semantic domain, while others maintain consistently low coefficients, reflecting more homogeneous, low-risk contexts.

As illustrated, even within the same semantic cluster, words exhibit heterogeneous patterns: some words have positive coefficients, indicating stronger associations with larger property losses, while others display negative coefficients, reflecting lower associated damage levels. This observation highlights that clusters capture thematic similarity, but do not necessarily imply uniform economic consequences. Hence, incorporating these regression coefficients into subsequent analyses provides an additional, economically grounded perspective. In particular, the integration of semantic clustering with property damage-based coefficients motivates the construction of a risk index that simultaneously reflects linguistic structure and financial severity, thereby offering a more comprehensive measure of fire-related risks.

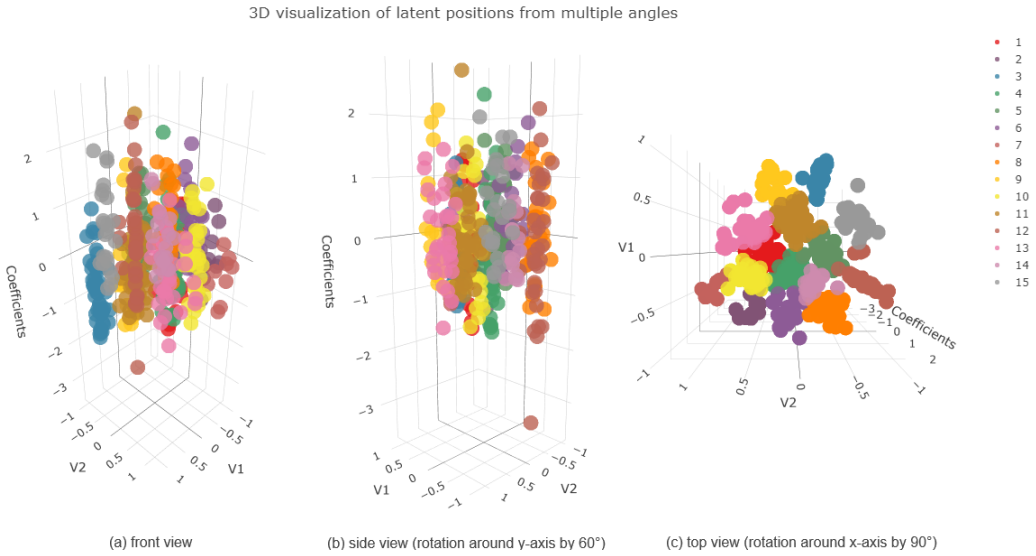


Figure 3. Three-dimensional visualization of latent word embeddings from multiple perspectives. The x- and y-axes represent latent positions, while the z-axis corresponds to coefficients estimated from the LASSO regression on fire-damage insurance claims. Panels (**a**), (**b**), and (**c**) show the front, side, and top views, respectively. Colors indicate clusters identified through k-means grouping.

7.3. Risk Index Estimation

The preceding analyses demonstrate that clusters derived from topic–word distributions provide semantically coherent categories of fire-related terms, while regression coefficients estimated from property damage amounts capture the associated economic severity. Importantly, these two perspectives highlight complementary aspects of risk: semantic clusters reflect the contextual mechanisms of fire occurrence, whereas coefficients quantify their financial impact. Moreover, as shown in the regression analysis, even words within the same cluster may exhibit heterogeneous patterns of association with damage amounts, indicating that semantic similarity alone is insufficient for fully characterizing fire risk.

To address this limitation, we propose the construction of a composite risk index that integrates both linguistic and economic dimensions. To quantify the relative contribution of words and clusters to fire-related incidents, we define three levels of risk indices: the word-level index $\gamma_{i,c}$, the cluster-level index δ_c, and the overall word index $\rho_{i,c}$. Each measure captures a distinct dimension of risk, ranging from fine-grained lexical associations to broader thematic categories. By jointly considering (i) cluster-level semantic associations and (ii) word-level coefficients derived from loss data, the risk index provides a systematic measure of fire risk that is interpretable in terms of language use and grounded in economic outcomes.

Word-Level Risk Index ($\gamma_{i,c}$)

As described in Step 4 of Figure 1, the Lasso-based risk modeling was applied to estimate the $\gamma_{i,c}$ that measures the relative contribution of word i within cluster c to the estimation of property damage amounts. A higher value indicates that the word is more strongly associated with larger expected losses compared to other words in the same cluster. Table 5 presents the top ten words with the highest $\gamma_{i,c}$ values in each cluster, along with their Risk Index. For example, in Cluster 2 (*Ignition from heat accumulation near combustible interior materials*) and Cluster 7 (*Ignition of flammable substances by friction heat or spark from machinery*), words such as *flame*, *heat*, and *grinder* show high $\gamma_{i,c}$ values, reflecting their strong linkage with higher levels of estimated property damage. Thus, $\gamma_{i,c}$ highlights words whose relative importance provides insight into the financial risk implications captured within each cluster.

Table 5. Cluster–word table with Risk Index values. For each cluster, the top ten words with the highest word-level Risk Index ($\gamma_{i,c}$) are reported. These words represent the relatively high-scoring lexical elements within each cluster, indicating stronger contributions to the estimated property damage amounts.

Cluster	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
1	photoinitiator (1.000)	ethyl acetate (0.911)	reactor (0.846)	heptane (0.814)	sclerotic (0.809)	centrifuge (0.798)	lesk (0.781)	cleaning solvent (0.741)	solvent (0.739)	chemical reaction (0.736)
2	flame (1.000)	plenty of (0.937)	wall (0.893)	interior wall (0.891)	laboratory (0.858)	floor (0.856)	radiant heat (0.834)	laboratory (0.736)	insulator (0.712)	exterior wall (0.674)
3	dust collection equipment (1.000)	urethane (0.907)	smoldering ignition (0.740)	rubber (0.714)	interior material (0.657)	heat (0.493)	cutting machine (0.445)	waste wood (0.082)	incinerate (0.000)	– (1.000)
4	stacked goods (0.987)	oil vapor (0.970)	storage tank (0.675)	gas (0.599)	oxygen (0.581)	reagent (0.448)	dust collector (0.427)	waste (0.381)	dummy (0.317)	oil (0.296)
5	manhole (1.000)	expander (0.949)	cover (0.939)	film (0.890)	epoxy (0.819)	mixer (0.801)	manufacturing building (0.798)	static electricity (0.683)	impurities (0.633)	physic (0.594)
6	shipping area (1.000)	petroleum product (0.879)	condenser (0.878)	storage (0.745)	cable (0.708)	refrigerator (0.691)	arson (0.642)	crayon (0.539)	oil tank (0.484)	hazardous substances plant (0.461)
7	grinder (1.000)	polystyrene (0.952)	fabric (0.909)	long time (0.902)	coating (0.816)	demolition (0.780)	manufacturing room (0.769)	panel (0.747)	electrical distribution (0.648)	cutting oil (0.637)
8	chemical material (1.000)	hazardous material (0.852)	leak (0.753)	vapor (0.680)	metal plate (0.603)	base material (0.544)	nucleic acid (0.456)	spontaneous combustion (0.424)	grease (0.402)	large-scale (0.341)
9	warehouse (1.000)	pallet (0.933)	storage (0.851)	cosmetic (0.825)	agitator (0.789)	drying room (0.712)	liquid (0.696)	secondary battery (0.693)	staff lounge (0.665)	heater rod (0.522)
10	activated carbon (1.000)	welding (0.863)	remain (0.816)	moisture (0.588)	semi-finished product (0.572)	accumulation (0.395)	tire (0.393)	paint (0.378)	plastic (0.369)	ignition (0.367)
11	service line (1.000)	packing room (0.869)	analysis lab (0.841)	interlayer short circuit (0.833)	distribution board (0.826)	molding machine (0.712)	unidentified short circuit (0.673)	power line (0.656)	trip (0.618)	air-compressor (0.616)

Table 5. *Cont.*

Cluster	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
12	automation (1.000)	flammable (0.938)	repair (0.843)	machine room (0.607)	insulation deterioration (0.541)	air-conditioner (0.356)	arc (0.328)	terminal (0.316)	battery (0.268)	outdoor unit of air-conditioner (0.258)
13	vulcanizer (1.000)	prefabricated (0.662)	research building (0.553)	indoor wiring (0.500)	dilution (0.492)	complete product (0.482)	automobile parts (0.480)	filtering equipment (0.465)	boiler (0.460)	circuit breaker (0.455)
14	roof (1.000)	entrance (0.969)	unidentified cause (0.889)	machine (0.885)	sandwich panel (0.829)	appliances (0.768)	dormitory (0.746)	compost (0.734)	boiler room (0.634)	deodorizing tower (0.609)
15	lower (1.000)	explosion (0.769)	electrical equipment (0.684)	commissioning (0.622)	duct (0.621)	rigid polyurethane (0.605)	stacked materials (0.595)	melting equipment (0.574)	condensation (0.562)	freeze drier (0.548)

Cluster-Level Risk Index (δ_c)

The index δ_c summarizes the risk associated with cluster c as a whole, aggregating the estimated coefficients of its constituent words. A higher δ_c indicates that, on average, words belonging to this cluster are strongly predictive of higher property losses. Table 6 reports these values. The five clusters with the highest δ_c include the following: (i) *Ignition of flammable vapor during organic solvent use*, (ii) *Electrical fires in areas handling flammable materials*, (iii) *Ignition from temperature rise in oil vapor areas*, (iv) *Ignition from heat accumulation near combustible interior materials*, and (v) *Fires or explosions from abnormal reactions during chemical handling*. These topics correspond to scenarios where ignition sources and flammable environments directly translate into severe financial consequences. By contrast, the clusters with the lowest δ_c, such as *Ignition from sparks inside dust collection equipment with filters and debris* or *Fires from electrical factors in process equipment sites*, represent situations with relatively weaker association to large-scale losses. In this way, δ_c provides an interpretable measure of how strongly each cluster of fire-related factors contributes to financial risk.

Table 6. Cluster ranking by Risk Index values and their associated topics. The Risk Index (δ_c) represents the aggregated risk contribution of each cluster, with higher values indicating strong fire-related financial losses.

Cluster	Risk Index	Topic
8	1.000	Ignition of flammable vapor during organic solvent use
6	0.954	Electrical fires in areas handling flammable materials
9	0.731	Ignition from temperature rise in oil vapor areas
2	0.639	Ignition from heat accumulation near combustible interior materials
5	0.626	Fires or explosions from abnormal reactions during chemical handling
12	0.606	Fires during machinery maintenance
7	0.561	Ignition of flammable substances by friction heat or spark from machinery
4	0.534	Spontaneous combustion from abandoned chemicals and oils
10	0.405	Fire from ignition in accumulated combustibles
13	0.388	Ignition in areas with accumulated combustible dust and oil vapors
14	0.344	Spontaneous combustion from improper oil residues storage or disposal
3	0.294	Ignition due to residual heat post-work with combustibles
1	0.216	Fire or explosion caused by chemical leakage
15	0.205	Ignition from sparks inside dust collection equipment with filters and debris
11	0.000	Fires from electrical factors in process equipment sites

Overall Word Risk Index ($\rho_{i,c}$)

The index $\rho_{i,c}$ extends beyond cluster membership to evaluate the global risk contribution of word i, accounting for its position across the latent embedding space. Table 7 lists the top twenty words by $\rho_{i,c}$. For example, *chemical material*, *shipping area*, and *hazardous material* emerge as the top three words. These terms directly connect to concrete accident scenarios, such as the generation of flammable vapors from sludge leakage, electrical fires in distribution boards, and vapor leakage during hazardous material processing. The $\rho_{i,c}$ index therefore highlights words that not only carry lexical salience within clusters, but also broader cross-cluster risk relevance.

Collectively, these three indices provide a multi-layered framework: $\gamma_{i,c}$ identifies salient words within clusters, δ_c ranks the clusters by their aggregate hazard potential, and $\rho_{i,c}$ detects globally critical words linked to real-world accident narratives.

Table 7. Word importance based on overall Risk Index ($\rho_{i,c}$). The table reports the top 20 words with the highest $\rho_{i,c}$ values, highlighting globally influential keywords strongly linked to insurance loss outcomes.

No.	Word	Risk Factor
1	chemical material	1.000
2	shipping area	0.977
3	hazardous material	0.926
4	petroleum product	0.917
5	condenser	0.916
6	warehouse	0.866
7	storage facility	0.850
8	pallet	0.832
9	cable	0.831
10	flame	0.820
11	manhole	0.813
12	automation	0.803
13	storage	0.791
14	plenty of	0.788
15	expander	0.788
16	cosmetic	0.778
17	stacked goods	0.761
18	agitator	0.760
19	film	0.758
20	spontaneous combustion	0.712

8. Discussion

Wehmeier and Mitropetrosb [31] analyzed the causes of fires in chemical plants based on incidents at a chemical–pharmaceutical company and categorized them as follows: Self-ignition (22%), Hot running of moving parts (17%), Welding (15%), Electrostatic (14%), Drying (10%), Repair/Maintenance (8%), Leakage (7%), and Electric (short-circuit) (7%). However, they also concluded that fire investigations in the German chemical industry reveal complex and heterogeneous causal structures. Furthermore, because this analysis relied solely on accident-frequency statistics, it was insufficient for assessing *risk* in the sense of loss severity. Complementarily, Darbra et al. [32] constructed a relative-probability event tree for major chemical accidents with domino effects, analyzing the causes, materials involved, effects and consequences, affected population, and the likelihood of specific accident sequences, thereby emphasizing the need to evaluate interconnected hazards rather than isolated causes.

To address these limitations, we replace frequency-based tallies with a composite, loss-aware scoring framework that links text-derived indicators to the scale of damage. The procedure is twofold: (i) estimate a risk score for each term that reflects its association with property-relevant losses, and (ii) interpret high-scoring terms by examining their local semantic neighborhoods to recover operational and environmental context.

We instantiate this framework via the Overall Word Risk Index, $\rho_{i,c}$, which quantifies the association between terms and loss outcomes rather than raw frequency. Using $\rho_{i,c}$ to rank and organize the vocabulary in the embedding (Figure 4), we identify four cross-cutting facets of fire risk: (i) *chemical leakage/vapors* (red), (ii) *storage/warehouses* (blue), (iii) *equipment/electrical* (green), and (iv) *self-ignition* (pink). The latent coordinates in Figures 2 and 4 are derived from the same embedding space, while Figure 4 highlights representative high-risk keywords and their corresponding topics to provide a more detailed interpretation across the four facets. In what follows, we analyze each facet and its representative cases.



Figure 4. Two-dimensional embedding (v1, v2) of fire-risk keywords, colored by four discussion facets and shaped by cluster assignments. Colored, labeled points are representative high-scoring terms (ranked by the Overall Word Risk Index $\rho_{i,c}$) in parentheses. Marker shapes encode cluster names, as shown in the legend; gray points denote the remaining vocabulary.

8.1. Four Cross-Cutting Facets of Fire Risk

First, the highest-ranking terms, such as chemical material, hazardous material, and petroleum product, capture risks associated with chemical leakage and the generation of flammable vapors. These findings align with well-documented hazards in chemical industries, where improper handling, leakage, or inadequate containment of chemical substances often lead to large-scale fire and explosion events. The 2009 Jaipur crude oil pipeline tank fire in India is an example of how a fire originating from a hazardous material leak can spread. The accident resulted in 13 deaths and over 200 injuries, setting a record for the worst accident of its kind in India [33]. The representative cases illustrate how unintended chemical leaks can escalate into severe incidents with significant property and human losses.

Second, structural and operational factors such as warehouses, storage facilities, and pallets are also prominent, underscoring the critical role of storage conditions and facility maintenance. Prior research on cold-chain logistics demonstrates that pallet stacking and package design strongly influence airflow and heat transfer within storage rooms, where insufficient ventilation can generate localized hotspots and elevate fire hazards [34]. In the context of energy storage, investigations highlight that inadequate cooling and airtight designs in portable energy storage systems promote overheating and exacerbate ignition risk [35]. Similarly, studies on lithium-ion batteries reveal that poor thermal management and accumulation of flammable gases in confined storage spaces can accelerate thermal runaway and combustion [36]. Together, these findings indicate that storage practices, ranging from palletized materials to battery warehouses, interact with flammable substances in ways that compound ignition potential, while recurring references to ventilation-related equipment emphasize how failures in airflow management can amplify both the likelihood and severity of fire incidents.

Third, a notable set of keywords points to ignition sources related to malfunctions or inadequate maintenance of electrical and mechanical equipment, such as condenser, cable, automation. These emphasize that not only causes directly related to chemical processes, but also the integrity of auxiliary mechanical and electrical systems play a crucial role in fire risk. Electrical fires are among the most universal causes of fire. Korean statistics show that 27.4% of fires incidents from 1996 to 2021 were attributed to electrical factors [37],

suggesting that the chemical industry is no exception to this risk. These findings suggest that the damage resulting from contact between common ignition sources and flammable raw materials or products handled in chemical plants can be significantly greater. The cases show that malfunctioning condensers, short-circuited cables, or maintenance using gas torches can trigger fires even in otherwise controlled environments.

Finally, terms such as spontaneous combustion and stacked goods reflect risks arising from self-ignition processes of chemicals and oil residues produced or derived in chemical plants, as well as environmental conditions. Recent studies on biomass storage have shown that moisture exchange, oxygen penetration, and heat accumulation can interact to induce self-heating and ultimately spontaneous ignition [38]. Similarly, investigations in mine waste dumps highlight that spontaneous combustion can compromise geomechanical stability, illustrating how natural or reactive processes extend the scope of fire hazards beyond purely mechanical or chemical failures [39]. These findings underscore the importance of incorporating self-ignition phenomena into fire risk assessments for chemical facilities and storage environments.

Taken together, these findings demonstrate that high-risk keywords are not confined to a single category, but span across chemical substances, storage practices, equipment reliability, and environmental interactions. By transforming unstructured fire investigation narratives into analyzable data, our framework systematically extracted semantically related terms through latent topic and embedding models, contextualized them with representative fire cases, and linked them to property damage estimates. This integrative approach culminated in the development of a composite Risk Index, which quantifies the relative importance of text-derived indicators in relation to financial loss outcomes. The combination of statistical evidence with real-world loss information provides interpretability and practical relevance, showing how latent textual patterns can be operationalized into measurable safety metrics. In doing so, the proposed risk index offers a model-based tool for identifying actionable risk factors that bridge large-scale textual evidence with practical fire safety management.

8.2. Risk Index Perspective

As summarized in Table 4, the *chemical leakage/vapors* facet concentrates the highest $\rho_{i,c}$ values (e.g., *chemical material*, *hazardous material*, *petroleum product*), indicating a stronger linkage to larger financial losses in our corpus—consistent with evidence that leakage and vapor formation are prominent drivers of severe incidents in chemical industries [31,33]. By contrast, the *self-ignition* facet (e.g., *spontaneous combustion*, *stacked goods*) exhibits lower $\rho_{i,c}$ values relative to the chemical-leakage/vapors facet in our corpus. This attenuation is consistent with three data-plausible mechanisms supported by prior work. First, self-heating often proceeds as a long-duration, low-temperature smouldering process with comparatively low heat-release rates and slow spread, which increases the opportunity for intervention before very large property losses accrue [40–42]. Second, consistent with prior surveys of major incidents, very large losses in hydrocarbon processing frequently originate from sustained leaks that evolve into flash fires or vapour cloud explosions, rather than from long-duration smouldering scenarios [32,43,44]. Third, contexts in which spontaneous combustion becomes catastrophic (e.g., biomass or mine-waste stockpiles) are emphasized in the engineering literature, but are under-represented in our chemical special-building corpus, attenuating the empirical linkage between self-ignition terms and large losses [38,39]. Interpreted this way, the relatively lower scores reflect dataset composition and event progression characteristics, rather than any contradiction with the established self-heating mechanism.

8.3. Why Similarly Clustered Terms Admit Distinct Operational Readings

Although cluster membership (shapes in Figure 4) reflects lexical neighborhoods, nearby terms can encode different operational contexts that co-occur in narratives. For example, *storage* terms (blue) can sit beside *equipment/electrical* terms (green) because pallet stacking and enclosure geometry restrict ventilation and heat rejection [34], thereby increasing the efficacy of routine electrical faults as ignition sources; likewise, *self-ignition* terms (pink) may appear near *leakage/vapor* terms (red) when oxygen ingress and heat accumulation in stacked goods elevate vapor formation and ignition potential [38]. Conversely, process-centric terms (e.g., maintenance/automation) may be spatially offset when hot-work contexts (sparks, localized heating) are discussed apart from storage constraints, yet they remain semantically bridgeable whenever operations occur proximate to flammable inventories or confined airflow paths [35,36]. Interpreting the embedding jointly with the loss-aware index $\rho_{i,c}$ (Figure 4) discriminates semantically adjacent yet economically distinct patterns: terms that lie close in the map but carry higher $\rho_{i,c}$ mark contexts historically associated with larger property losses, whereas nearby low-$\rho_{i,c}$ terms indicate operational exposure without the same tail severity. This joint reads proximity for mechanism $\rho_{i,c}$, which yields a ranked set of actionable priorities, directing inspection toward leakage/vapor configurations while maintaining vigilance for storage and self-ignition scenarios that can escalate under adverse conditions.

The proposed framework not only enables post-incident classification of fire causes, but also provides a practical foundation for preventive fire management. By translating text-derived risk indicators into measurable patterns, the derived Risk Index helps identify high-risk operational contexts and materials before incidents occur. Table 8 illustrates how the ranked Risk Index can guide on-site inspections and safety checklists in industrial facilities. This approach supports decision-makers in prioritizing safety measures under limited manpower and budget, and establishes a systematic linkage between historical fire evidence and actionable preventive guidelines. Moreover, the framework can be extended by incorporating real-time sensor-based parameters such as temperature, dust concentration, or gas-leakage levels to enable proactive monitoring analogous to diagnostic systems in medicine. In practice, these parameters can be included as additional covariates in the Risk Index estimation stage, allowing for the model to quantify how environmental conditions influence fire risk levels. By jointly modeling textual indicators and measurable physical factors, the extended framework can provide more timely and context-aware assessments, supporting preventive decision-making in high-risk industrial settings.

Table 8. Example of fire prevention guidelines for industrial facilities by Risk Index ranking.

Risk Index	Cluster No.	Topic	Main Keywords	Related Words	Fire and Explosion Prevention Checklist
1.000	8	Ignition of Flammable Vapor during Organic Solvent Use	chemical material, hazardous material	leak, paint, sludge, grease, vapor, mixing equipment	• Are measures in place to prevent leakage of flammable chemical materials? • Is there any accumulation of sludge such as grease? • Is the ventilation system operating to prevent retention of flammable vapors? • Are vapors accumulating where sparks from mixing equipment could ignite?
0.977	6	Electrical Fires in Areas Handling Flammable Materials	shipping area, distribution board	–	• Are wiring systems, including distribution boards in the shipping area, properly maintained?

Table 8. *Cont.*

Risk Index	Cluster No.	Topic	Main Keywords	Related Words	Fire and Explosion Prevention Checklist
0.866	9	Ignition from Temperature Rise in Oil Vapor Areas	warehouse, breakdown, discharge, storage, secondary battery, synthetic resins, hot work equipment	secondary battery, discharge, hot work equipment, synthetic resins	• Are discharged secondary batteries stored safely? • Is hot work inside the warehouse performed under safe conditions? • Are combustible materials stored at a safe distance from hot work areas?

9. Conclusions

This study presents a data-driven framework for quantifying and interpreting fire-related risk factors embedded in unstructured investigation narratives. By transforming textual records into latent topical structures and constructing an embedding-based network of word interactions, we identified semantically coherent clusters that capture operational and environmental contexts of fire incidents. The key contribution of this work lies in linking these text-derived semantic signals with structured financial loss data to develop a composite metric, the Overall Word Risk Index ($\rho_{i,c}$), which highlights the words and topics most closely associated with real-world damage outcomes. The resulting framework provides an interpretable and loss-aware map of industrial fire risk, offering a foundation for evidence-based prevention and safety management.

In this study, the derived Risk Index can be considered useful as reference data for safety management and insurance underwriting of industrial sites. Within a workplace, the Risk Index can help assign priorities when establishing preventive measures against potential accidents inferred from topics and words. Given the limited manpower and budget typically available at industrial sites, allocating safety management resources in proportion to the degree of risk is of critical importance. In the context of insurance subscription, understanding the level of risk at a workplace is also of utmost priority. Although insurance premiums are generally determined based on statistical loss ratios by industrial sector, it remains difficult to assess the unique risks arising from a site's specific operating conditions. In this regard, the Risk Index serves as a quantitative measure of a facility's risk level and may provide valuable guidance for insurance underwriters when determining whether to underwrite facilities that involve substances or processes associated with higher risk scores.

Although the results are promising, potential threats to validity should also be acknowledged. The narratives used in this study were prepared by professional fire officers based on verified evidence and standardized terminology, ensuring a high degree of internal validity with minimal subjective bias. While supervised validation would ideally complement our approach, the current unsupervised framework incorporates expert interpretation to maintain analytical reliability. Regarding external validity, the proposed methodology could be extended to other domains such as manufacturing, logistics, construction, or residential and public facilities to assess its generalizability and adaptability across diverse fire contexts.

Future research will extend the current framework in two directions. First, we plan to develop methods for systematically tracking the temporal trends of risk factors, enabling the identification of how the prevalence and severity of specific hazards evolve over time. Such analyses will provide an evidence-based foundation for monitoring emerging risks and for designing timely interventions. Second, we aim to advance causal inference approaches tailored to text-derived risk indicators, with a particular focus on the risky keywords identified in investigation records. By establishing causal relationships rather

than mere associations, this line of work will enhance the interpretability of the extracted factors and strengthen their utility for policy and decision-making in fire risk management.

Author Contributions: Conceptualization, H.J., S.A. and Y.J.; methodology, Y.J.; software, Y.J.; validation, H.J. and S.A.; formal analysis, Y.J.; investigation, H.J.; resources, H.J. and S.A.; data curation, H.J. and S.A.; writing—original draft preparation, H.J. and Y.J.; writing—review and editing, Y.J.; visualization, H.J. and Y.J.; supervision, Y.J.; project administration, H.J., S.A. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study were obtained from Korea Fire Protection Association under license, and restrictions apply to the availability of these data. Data are available from the authors upon reasonable request and with the permission of Korea Fire Protection Association.

Acknowledgments: The authors thank the Korean Fire Protection Association for providing access to fire investigation records, without which this study would not have been possible.

Conflicts of Interest: Two authors (H.J. and S.A.) are affiliated with the Korean Fire Protection Association, which provided access to the fire investigation records analyzed in this study. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The authors declare no other conflicts of interest.

References

1. Kim, B.; Hwang, K.I. Text mining techniques to identify causes and hazards of ship fire accidents. *J. Mar. Eng. Technol.* **2020**, *44*, 189–195. [CrossRef]
2. Tirunagari, S. Data mining of causal relations from text: Analysing maritime accident investigation reports. *arXiv* **2015**, arXiv:1507.02447. [CrossRef]
3. Liu, X.; Wang, J.; Zhang, T. Text Mining Approach for Trend Tracking in Scientific Research: A Case Study on Forest Fire. *Fire* **2023**, *6*, 33. [CrossRef]
4. Zhao, Q.; Huang, T.; Wang, S. Assessing Trends in Wildland-Urban Interface Fire Research Through Text Mining: A Comprehensive Analysis of Published Literature. *J. For. Res.* **2024**, *35*, 71. [CrossRef]
5. Zhang, Y.; Chen, H.; Liu, W. Risks Identification and Fire Scenarios Determination of Ship Fires Based on Improved Text Mining and K-Means Algorithm. *Fire Mater.* **2025**, *49*, 904–917.
6. Wang, L.; Xu, Z.; Li, J. Holistic Value-at-Risk Assessment Framework for Fire Risk Assessment of Heritage Buildings Based on Analytic Hierarchy Process and Text Mining. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng.* **2023**. [CrossRef]
7. Wang, Y.; Guo, D.; Liu, P. A New Text-Mining–Bayesian Network Approach for Identifying Chemical Safety Risk Factors. *Mathematics* **2022**, *10*, 4815. [CrossRef]
8. Li, J.; Zhou, Y.; Chen, Z. Identifying Coal Mine Safety Production Risk Factors by Employing Text Mining and Bayesian Network Techniques. *Process Saf. Environ. Prot.* **2022**, *162*, 1067–1081. [CrossRef]
9. Liu, Z.; Wang, Y.; Lee, H. Cause Analysis of Hot Work Accidents Based on Text Mining and Deep Learning. *J. Loss Prev. Process Ind.* **2022**, *76*, 104747.
10. Yang, C.; Zhang, F.; Hu, L. Analysis of Risk Factors of Coal Chemical Enterprises Based on Text Mining. *J. Environ. Public Health* **2023**, *2023*, 4181159. [CrossRef]
11. Chen, R.; Liu, X.; Zhao, M. A Chemical Accident Cause Text Mining Method Based on Improved Accident Triangle. *BMC Public Health* **2024**, *24*, 39. [CrossRef] [PubMed]
12. Russo, M.; Fischer, A.P.; Huber-Stearns, H.R. Wildfire narratives: Identifying and characterizing multiple understandings of western wildfire challenges. *Environ. Sci. Policy* **2024**, *160*, 103824. [CrossRef]
13. Ma, C.; Du, H.; Luan, S.; Dong, E.; Gardner, L.M.; Gernay, T. From Occurrence to Consequence: A Comprehensive Data-driven Analysis of Building Fire Risk. *arXiv* **2025**, arXiv:2503.22689.
14. Zhang, S.; Song, Y.; Dong, Q.; Yang, H.; Shi, L. Developing an indicator system for urban fire risk assessment. *Fire Saf. J.* **2025**, *158*, 104536. [CrossRef]

15. Dapeng, G. Higher Emissions Scenarios Increase Wildland–Urban Interface Fire Hazard in China. *Sustainability* **2025**, *17*, 7409. [CrossRef]
16. Kudryavtsev, S.; Yemelin, P.; Rakhimberlina, A.; Derbush, S.; Kinayatov, M. Modern Approaches to Fire Safety Risk Assessment at Chemical Industry Enterprises. *Tr. Univ.* **2025**, *1*, 93–104. [CrossRef]
17. Kim, S.; Park, J.; Lee, K. Experimental Study on Short-Text Clustering Using Transformer-Based Semantic Similarity Measure. *PeerJ Comput. Sci.* **2024**, *10*, e2078. [CrossRef]
18. Park, Y.; Choi, D. Investigating Transfer Learning Capacity of SBERT Models in Text Clustering. *Appl. Comput. Rev.* 2024, *in press*.
19. Choi, M.; Lee, E. A Novel Text Clustering Model Based on Topic Modelling and Autoencoders. *Chaos Solitons Fractals* **2024**, *181*, 096007792400184X.
20. Lee, J.; Kim, T.; Park, S. Text Clustering Based on Pre-Trained Models and Autoencoders. *Front. Comput. Neurosci.* **2024**, *11*, 1334436.
21. Stefanovitch, N.; Jacquet, G.; De Longueville, B. Graph and Embedding based Approach for Text Clustering: Topic Detection in a Large Multilingual Public Consultation. In Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), Austin, TX, USA, 30 April–4 May 2023; ACM: New York, NY, USA, 2023. [CrossRef]
22. Jeong, H.; Park, D. Network Analysis of Social Media Texts. *Front. Res. Metr. Anal.* **2025**, *10*, 1558798. [CrossRef]
23. Suurmond, R.; Wynstra, F.; Vermeij, A.; Haag, E.J. Text mining and network analytics for literature reviews: Exploring the landscape of purchasing and supply management research. *J. Purch. Supply Mgmt.* **2024**, *30*, 100892. [CrossRef]
24. Anderson, M.; Jacobs, B. *Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns*; Jacobs Levy Equity Management Center for Quantitative Financial Research Paper; Jacobs Levy Equity Management Center for Quantitative Financial Research: Philadelphia, PA, USA, 2023.
25. Rahman, A.; Lee, N. Exploring Advanced Techniques in Natural Language Processing and Machine Learning for In-depth Analysis of Insurance Claims. In *Smart Computing Paradigms: Artificial Intelligence and Network Applications*; Springer: Singapore, 2024; pp. 47–56.
26. Jeon, Y.; Park, J.; Jin, I.H.; Chung, D. Network-based topic structure visualization. *J. Appl. Stat.* **2025**, *52*, 509–523. [CrossRef]
27. Yan, X.; Guo, J.; Lan, Y.; Cheng, X. A biterm topic model for short texts. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 1445–1456.
28. Jeon, M.; Jin, I.; Schweinberger, M.; Baugh, S. Mapping Unobserved Item–Respondent Interactions: A Latent Space Item Response Model with Interaction Map. *Psychometrika* **2021**, *86*, 378–403. [CrossRef] [PubMed]
29. Liu, J.S. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.* **1994**, *89*, 958–966. [CrossRef]
30. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
31. Wehmeier, G.; Mitropetrosb, K. Fire protection in the chemical industry. *Chem. Eng.* **2016**, *48*, 259–264.
32. Darbra, R.; Palacios, A.; Casal, J. Domino effect in chemical accidents: Main features and accident sequences. *J. Hazard. Mater.* **2010**, *183*, 565–573. [CrossRef]
33. Lee, J.; Ryu, J.; Park, S.; Yoon, M.O.; Lee, C. Study on the Evaluation of Radiant Heat Effects of Oil Storage Tank Fires Due to Environmental Conditions. *Fire Sci. Eng.* **2020**, *34*, 72–78. [CrossRef]
34. Ahmad, N.E.; Duret, S.; Moureh, J. Interactions between package design, airflow, heat and mass transfer, and logistics in cold chain facilities for horticultural products. *Energies* **2022**, *15*, 8659. [CrossRef]
35. Eslami Majd, A.; Tchuenbou-Magaia, F.; Meless, A.M.; Adebayo, D.S.; Ekere, N.N. A review on cooling systems for portable energy storage units. *Energies* **2023**, *16*, 6525. [CrossRef]
36. Feng, X.; Ouyang, M.; Liu, X.; Lu, L.; Xia, Y.; He, X. Thermal runaway mechanism of lithium ion battery for electric vehicles: A review. *Energy Storage Mater.* **2018**, *10*, 246–267. [CrossRef]
37. Lee, H.G.; Son, U.N.; Je, S.M.; Huh, J.H.; Lee, J.H. Overview of fire prevention technologies by cause of fire: Selection of causes based on fire statistics in the Republic of Korea. *Processes* **2023**, *11*, 244. [CrossRef]
38. Wei, J.; Yao, C.; Sheng, C. Modelling self-heating and self-ignition processes during biomass storage. *Energies* **2023**, *16*, 4048. [CrossRef]
39. Nguyen, P.M.V. A Review of the Impact of Spontaneous Combustion on Slope Stability in Coal Mine Waste Dumps. *Appl. Sci.* **2025**, *15*, 7138. [CrossRef]
40. Rein, G. Smouldering Combustion Phenomena in Science and Technology. *Int. Rev. Chem. Eng.* **2009**, *1*, 3–18.
41. Torero, J.L.; Gerhard, J.I.; Martins, M.F.; Zanoni, M.A.B.; Rashwan, T.L.; Brown, J.K. Processes Defining Smouldering Combustion: Integrated Review and Synthesis. *Prog. Energy Combust. Sci.* **2020**, *81*, 100869. [CrossRef]
42. Santoso, M.A.; Christensen, E.G.; Yang, J.; Rein, G. Review of the Transition From Smouldering to Flaming Combustion in Wildfires. *Front. Mech. Eng.* **2019**, *5*, 49. [CrossRef]

43. Atkinson, G.; Cowpe, E.; Halliday, J.; Painter, D. A review of very large vapour cloud explosions: Cloud formation and explosion severity. *J. Loss Prev. Process Ind.* **2017**, *48*, 367–375. [CrossRef]
44. Chang, J.I.; Lin, C.C. A study of storage tank accidents. *J. Loss Prev. Process Ind.* **2006**, *19*, 51–59. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.