Computing, Telecommunication and Control, 2026, Vol. 19, No. 1, Pp. 91–102.
Информатика, телекоммуникации и управление. 2026. Том 19, № 1. С. 91–102.

Research article
DOI: https://doi.org/10.18721/JCSTCS.19109
UDC 004.91:005.334

METHOD FOR CLASSIFYING RISK INCIDENTS BASED ON SELF-ORGANIZATION OF SEMANTIC CLUSTERS

V.P. Shkodyrev, D.G. Rodionov , E.A. Konnikov ✉

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ konnikov.evgeniy@gmail.com

Abstract. A method for automatic classification of textual descriptions of emergency risk incidents based on self-organizing semantic clustering is presented, which does not require prior data labeling. Unlike traditional approaches, the method involves a two-stage scheme, which consists of self-organization of a latent taxonomy of incidents through hierarchical thematic decomposition of the text corpus, as well as continuous classification of new messages according to their degree of belonging to all automatically selected classes at once. This transition from rigid assignment to a single class to fuzzy membership allows hybrid incidents to be decomposed into several risk factors, reflecting their mixed nature. The developed algorithm forms an interpretable and stable taxonomy of incidents that preserves the structural isolation of clusters even with a high proportion of hybrid events. Testing on the NRC data corpus showed that most messages have a dominant risk factor with significant secondary components. The average semantic consistency of clusters was ~0.62 (cosine measure), and the classification confidence is distributed around the mean, reflecting the presence of both pure and mixed incidents. The results confirm that the proposed method provides a mathematically correct decomposition of complex situations into a set of risk factors and reduces the sensitivity of classification to noise and inaccuracies in the input text. The methodology is focused on proactive risk analysis in complex technical systems and can be used for automated decision support in industrial safety systems.

Keywords: risk incidents, unstructured data, semantic analysis, thematic modeling, clustering, fuzzy classification, risk taxonomy

Acknowledgements: The research was financially supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state assignment "Development of methodology for the formation of a tool base for analysis and modeling of spatial socio-economic development of systems in the conditions of digitalization with reliance on internal reserves" (FSEG-2023-0008).

Citation: Shkodyrev V.P., Rodionov D.G., Konnikov E.A. Method for classifying risk incidents based on self-organization of semantic clusters. Computing, Telecommunications and Control, 2026, Vol. 19, No. 1, Pp. 91–102. DOI: 10.18721/JCSTCS.19109

© Shkodyrev V.P., Rodionov D.G., Konnikov E.A., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19109
УДК 004.91:005.334

МЕТОД КЛАССИФИКАЦИИ РИСК-ИНЦИДЕНТОВ НА ОСНОВЕ САМООРГАНИЗАЦИИ СЕМАНТИЧЕСКИХ КЛАСТЕРОВ

В.П. Шкодырев, Д.Г. Родионов , Е.А. Конников ✉

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ konnikov.evgeniy@gmail.com

Аннотация. Представлен метод автоматической классификации текстовых описаний аварийных риск-инцидентов на основе самоорганизующейся семантической кластеризации, не требующий априорной разметки данных. В отличие от традиционных подходов, метод предполагает двухэтапную схему, которая заключается в самоорганизации латентной таксономии инцидентов посредством иерархического тематического разложения текстового корпуса, а также непрерывной классификации новых сообщений по степени принадлежности ко всем автоматически выделенным классам сразу. Такой переход от жесткого присваивания одного класса к нечеткой принадлежности позволяет декомпозировать гибридные инциденты на несколько факторов риска, отражая их смешанную природу. Разработанный алгоритм формирует интерпретируемую и устойчивую таксономию инцидентов, сохраняющую структурную обособленность кластеров даже при высокой доле гибридных событий. В рамках апробации на корпусе данных NRC показано, что большинство сообщений имеют доминирующий фактор риска при наличии значимых вторичных компонентов. Средняя семантическая согласованность кластеров составила ~0.62 (косинусная мера), а уверенность классификации распределена вокруг среднего значения, отражая наличие как чистых, так и смешанных инцидентов. Результаты подтверждают, что предложенный метод обеспечивает математически корректную декомпозицию сложных ситуаций на совокупность факторов риска и снижает чувствительность классификации к шуму и неточностям входного текста. Методология ориентирована на проактивный анализ риска в сложных технических системах и может применяться для автоматизированной поддержки принятия решений в рамках систем промышленной безопасности.

Ключевые слова: риск-инциденты, неструктурированные данные, семантический анализ, тематическое моделирование, кластеризация, нечеткая классификация, таксономия рисков

Финансирование: Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания «Разработка методологии формирования инструментальной базы анализа и моделирования пространственного социально-экономического развития систем в условиях цифровизации с опорой на внутренние резервы» (FSEG-2023-0008).

Для цитирования: Shkodyrev V.P., Rodionov D.G., Konnikov E.A. Method for classifying risk incidents based on self-organization of semantic clusters // Computing, Telecommunications and Control. 2026. Т. 19, № 1. С. 91–102. DOI: 10.18721/JCSTCS.19109

Introduction

Ensuring industrial safety of complex technical systems requires systematic accounting of diverse operational information, including text reports on incidents and accidents. Analysis of such weakly structured data allows extracting hidden patterns of accidents and identifying risk factors for taking preventive measures. Traditional methods of probabilistic risk and reliability analysis rely on failure and event statistics presented in a strictly structured form [1, 2]. Classical models are difficult to apply

© Шкодырев В.П., Родионов Д.Г., Конников Е.А., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

directly to text descriptions that do not contain formalized markup of the causes and consequences of accidents. In real operating logs, a single incident report often combines several aspects: the cause, development, and outcome of the event. Rigidly assigning such a report to a single category leads to a loss of information about the mixed nature of the incident and artificially inflates the number of classes.

The problem of automated interpretation of weakly structured data in the security management loop is interdisciplinary in nature, combining information theory, cybernetics, linguistics, and reliability theory. The foundations of the infometric approach – the study of statistical patterns in information flows (messages) in complex systems – were laid as early as the mid-20th century. C. Shannon introduced quantitative measures of information and uncertainty, laying the foundation for communication theory [3]. N. Wiener and his followers developed a cybernetic approach to management and communication in technology and organisms [4]. The laws of information distribution were empirically discovered, in particular Zipf's and Lotka's laws, which describe the frequency distribution of terms and messages [5]. These classic works showed that even unstructured data contains hidden statistical structures.

The current stage is characterized by rapid development of natural language processing (NLP) and machine learning methods, which opens up new opportunities for analyzing text data in the field of industrial safety. Most known studies focus on the automatic classification of reports according to predetermined categories or risk attributes. For example, models based on image recognition methods and deep neural networks have been proposed, trained to assign incident descriptions to one of several fixed classes [6–8]. The practical application of such approaches has been demonstrated using data from the construction industry and other fields [9–11]. However, experts note that simple categorization is insufficient for a complete understanding of the causes and trends of accidents. A classifier that assigns a single label to a report does not reveal the internal relationships between risk factors and does not reflect the "hybrid" nature of incidents involving multiple cause-and-effect components.

An alternative to rigid classification is semantic text analysis, in which hidden topics or features are automatically extracted from the corpus. Classic methods here are latent semantic analysis and probabilistic topic modeling (Latent Dirichlet Allocation, LDA) [12]. Topical modeling has been successfully used to identify common accident and incident scenarios without manual annotation [13]. For example, in [14], the BERTopic thematic model (based on BERT contextual embeddings) was used to automatically identify key themes and trends in occupational injuries from 22623 OSHA reports. This approach made it possible to detect hidden accident scenarios and predisposing factors that were difficult to identify using keyword methods. The advantage of thematic models is their ability to analyze large amounts of text data and dynamically update the taxonomy as new information becomes available. However, this approach also has limitations. First, standard algorithms such as LDA do not work well on short technical records due to data sparsity. Second, many modern deep models are "black boxes" – they improve the metric representation of text but lose transparency and interpretability, which is unacceptable for critical system security tasks. In addition, topics obtained automatically do not have unambiguous names, requiring expert interpretation. Accordingly, thematic analysis of incident data must be supplemented with structuring and decomposition steps so that the extracted topics acquire the meaning of risk factors and can be used in assessing and mitigating that risk.

A separate area is related to the ontological approach and semantic-categorical analysis of texts based on specified thesauri. Systems have been created where expert-developed hazard ontologies are used to tag incident reports before they are analyzed [15]. Such solutions allow for highly accurate interpretation, but require a lot of effort to maintain the knowledge base and do not scale well to new types of threats. In dynamically changing operating conditions, this makes it difficult to use ontological systems for accident prediction. Thus, there is a contradiction. On the one hand, loosely structured text information needs to be taken into account to improve proactive risk management; on the other hand, there is no single set of methods capable of automatically converting raw text into a metric space of threats while preserving their structure, causal relationships, and the ability to interpretably track changes in risk.

In recent years, there has been an increasing number of studies devoted to the analysis of unstructured accident descriptions using NLP. Most of them address the task of classifying incidents by type or factor, which is similar to our approach. For example, one study proposes an approach based on gradient boosting and text processing methods for the automatic classification of construction accident records into predefined risk categories [16]. Another study developed a comprehensive machine learning framework for extracting hazardous factors from report texts and analyzing their trends over time [17]. OSHA reports were processed using classification and clustering methods, which made it possible to identify the growth of certain risks and vulnerable groups of workers. In other areas, various authors have applied the GPT-4 model (large language model) to automatically categorize textual descriptions of risks in construction [18–20]. It has been shown that modern Large Language Models (LLMs) are capable of achieving accuracy comparable to classical algorithms on structured features. However, LLM models require careful prompt engineering, their output is difficult to verify, and the use of commercial APIs in critical systems is undesirable for data security reasons. In addition, all of the above approaches involve discrete classification – assigning a single class to each message. This approach does not reflect the full scope of information if the incident is of a mixed nature.

Against this backdrop, an uncontrolled multi-class division of incident space with the possibility of soft, diverse classification of each event looks promising. Thematic modeling approaches indicate that such a taxonomy can self-organize directly from the data, without pre-assigning classes. The question of structuring the obtained themes into risk factors and classifying new messages according to all factors at once remains open. In our work, we have attempted to solve this problem. At the first stage, a method of contextual regularization of the feature space was developed, which eliminates the effect of sparsity and increases the connectivity of the semantic space of documents. Heterogeneous messages became geometrically comparable, forming the basis for the selection of taxonomy. Next, a scenario-optimization method for semantic processing of accident data was proposed, which allows ranking risks in the text according to the degree of controllability [21]. Nevertheless, these solutions left the problem of systematic classification of incidents unresolved. It is necessary to combine the identified latent themes into a stable taxonomy and describe each message with a profile of its belonging to this structure.

The goal of this work is to develop a teacherless risk incident classification method that allows for the automatic construction of a hierarchical semantic taxonomy of emergency situations and the fuzzy classification of each incident across a set of selected classes. The scientific novelty of the method lies in the transition from traditional discrete classification logic to continuous assessment of the degree to which an event belongs to all potential classes simultaneously. This ensures the correct decomposition of complex, combined incidents into constituent risk factors and reduces the sensitivity of classification to local changes in the message text.

Methods

The proposed method treats the classification task as a two-stage process, which is represented in the block diagram below. In the first stage, a taxonomy – a set of semantic classes without markup – is automatically self-organized from the corpus of incident descriptions. This is achieved through hierarchical thematic modeling of the latent feature space. In the second stage, new events are continuously marked up relative to the resulting taxonomy – each message is assigned not one tag, but a vector of memberships across all classes at once.

Let there be a corpus of N documents (incident descriptions) represented as non-negative feature vectors $x_d \in R^M$ dimension M. All documents form a matrix of characteristics $X = \left[x_{dm} \right] \in R^{N \times M}$.

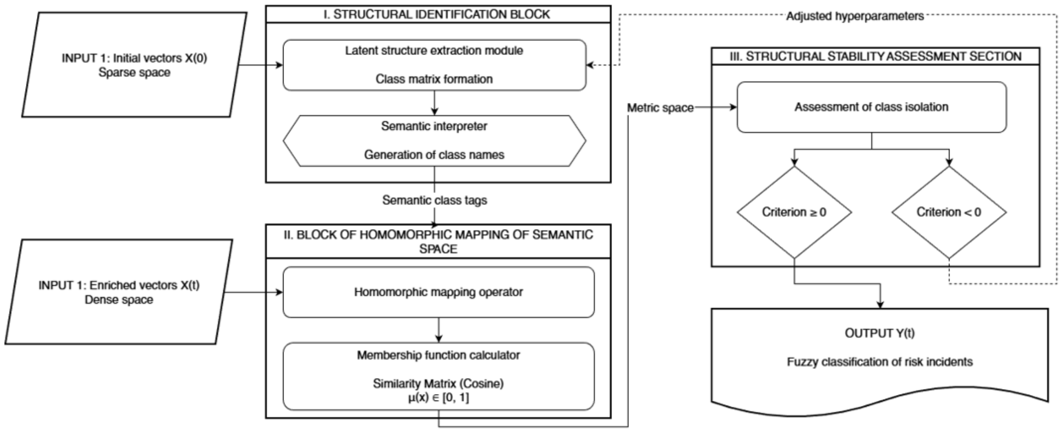

Fig. 1. Method for classifying risk incidents based on self-organization of semantic clusters and soft labeling

To build a semantic taxonomy, the number of large thematic clusters is specified at the top level K. A probabilistic model of thematic decomposition is applied, resulting in the calculation of a matrix of document distribution by topic $\Theta = [\theta_{dk}] \in R^{N \times K}$, where $\theta_{dk} = P(z = k | d)$ – posterior probability (weight) of topic k in document d, matrix of topic profiles $B = [\beta_{km}] \in R^{K \times M}$, where $\beta_{km} = P(w = m | z = k)$ – probabilities of terms in the topic. Thus, at the first level, each document is characterized by a vector $\theta_d = (\theta_{d1}, \ldots, \theta_{dK})$. To fix the discrete taxonomy, we perform a strict assignment of the document to one of the upper classes according to the Maximum A Posteriori (MAP) rule:

$$c_1(d) = \arg\max_{1 \le k \le K} \theta_{dk}.$$

In other words, the document relates to the topic that carries the most weight θ_{dk}. After that, subtopics are locally identified within each obtained top cluster k. For each k the number of subclusters is specified L, and by subset of documents $d : c_1(d) = k$ a thematic model with L themes is being built again. Thus, local distributions are formed $\Theta^{(k)} = \left[\theta^{(k)} dj\right]$ and profile matrices $B^{(k)} = \left[\beta^{(k)}\right]$ for $j = 1 \ldots L$ inside class k. Document d, related to the top topic k, at the second level is characterized by the vector $\theta_d^{(k)} = \left(\theta_{d1}^{(k)}, \ldots, \theta_{dL}^{(k)}\right)$ by subtopics of this topic. Applying a similar MAP rule, we fix the sheet cluster:

$$c_2(d) = \arg\max_{1 \le j \le L} \theta_{dj}^{(k)}, \quad k = c_1(d).$$

The pair (k, j) defines the final leaf cluster of the taxonomy: $C_{k,j} = d : c_1(d) = k,\ c_2(d) = j$. Thus, taxonomy $\mathcal{C} = C_{k,j}$ is extracted from the data structure rather than from predefined expert classes. Each cluster $C_{k,j}$ corresponds to a certain set of incidents that are similar in meaning, selected automatically.

In the next step of the developed method, discrete taxonomy is transformed into a continuous classification model. The idea is not to limit oneself to a single label $C_{k,j}$ to the document, but to calculate the degrees of belonging of the document to all clusters at once. To do this, each cluster is assigned a short semantic name $l_{k,j}$. The set of all tags forms a set $L = l_{k,j}$ in size $K! \times !L$. Then, a common metric space of features is constructed, in which documents and cluster labels are represented as embedding vectors of unit length. For this purpose, one can use, for example, normalized vectors of rows of matrix X after contextual regularization. They contain distributed semantic representations of documents. Labels $l_{k,j}$ can be represented in the same vector space, for example, by their characteristic words. Let us denote by u_d the normalized vector of document d, and through u_l – normalized label vector $l \in L$. Then, naturally, a fuzzy membership matrix is introduced $F = [f_{d,l}]$ dimensions $N \times |L|$, whose elements are calculated as the cosine similarity between the document and the cluster label:

$$f_{d,l} = \cos\left(u_d, u_l\right) \in \left[0,1\right].$$

Each row of the matrix F, $f_d = \left(f_{d,1}, \ldots, f_{d,|L|}\right)$, is a continuous profile of document d's membership in all taxonomy classes simultaneously. If necessary, this profile can be interpreted probabilistically by performing normalization. Let us introduce a normalized measure:

$$p_{d,l} = \frac{\left(f_{d,l}\right)^{\lambda}}{\sum_{l' \in L}\left(f_{d,l'}\right)^{\lambda}},$$

where $\lambda > 0$ is the parameter that controls the "stiffness" of the distribution. The resulting vector $p_d = \left(p_{d,l}\right) l \in L$ is a generalized (fuzzy) classification of incident d. If the text contains several independent semantic components, their corresponding values p will be relatively large, and the document will be described by a mixture of classes rather than a single label.

The transition from a discrete label to a membership vector is fundamentally important not only conceptually, but also mathematically. Function $f_{d,l}$ proves to be resistant to minor disturbances in the semantic representation of the document. Indeed, consider a fixed cluster l and two documents i and j. For the difference between their attributes, the following inequality holds true:

$$\left|f_{i,l} - f_{j,l}\right| = \left|\cos\left(u_i, u_l\right) - \cos\left(u_j, u_l\right)\right| \le \left|u_i - u_j\right|.$$

Since the vectors are normalized, a small change in the text results in a small change in the embedding u_i. This leads only to a slight change in the overall profile. f_i, whereas classical discrete classification could change the label abruptly. Thus, the transition from single labels to continuous membership scores not only enables multi-class decomposition, but also reduces the model's sensitivity to noise, which is inevitable in text data. This is particularly relevant for technical journals, where incident descriptions may contain inaccuracies and variable wording.

Results

The developed method was tested on a corpus of U.S. Nuclear Regulatory Commission (NRC) reports on operational events at nuclear power plants. $N = 10000$ text records were used, covering a wide range of incidents, from equipment failures to human errors and natural phenomena. A standard NLP pipeline was used to convert the texts into matrix form X: cleaning, tokenization, lemmatization of terms, and TF–IDF weighting. Next, contextual regularization of matrix X was performed, namely, the construction of a graph of joint term co-occurrence followed by diffusion reordering of features. As a result, the feature space dimension was $M \approx 5000$, and the semantic coherence of the data increased significantly. The taxonomy parameters were selected structurally. $K = 5$ top-level topics and $L = 3$ subtopics were set for each, resulting in $K \times L = 15$ final clusters. These values are consistent with expert opinion – the corpus is expected to contain about a dozen broad categories of incidents.

The figure below shows the distribution of classification confidence values $\xi_d = \max_{l \in L} p_{d,l}$ according to all documents. The amount ξ_d reflects how clearly the incident is associated with a single

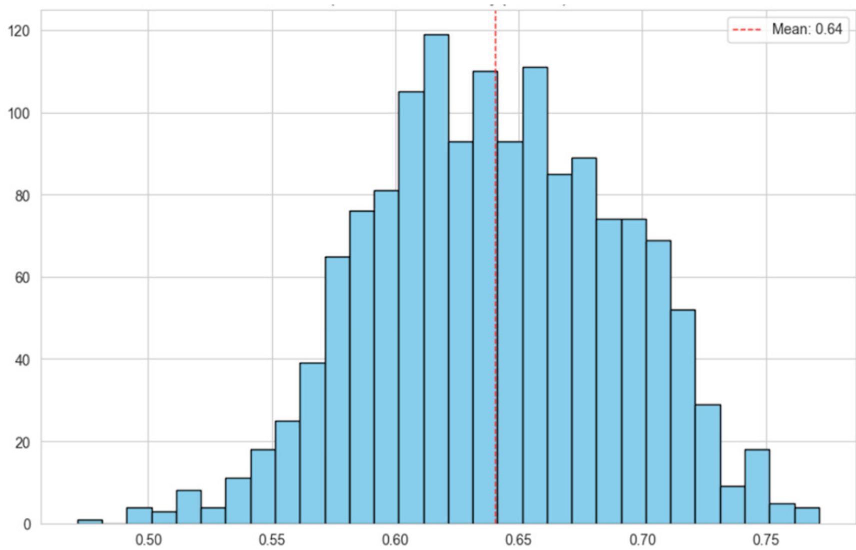

Fig. 2. Distribution of soft classification confidence

category in the model. When $\xi_d = 1$ the document strictly belongs to one cluster, with smaller ξ_d several classes have comparable weights. The histogram shows the concentration of values around the mean $\overline{\xi} \approx 0.64$. This means that for most events, the model identifies a pronounced dominant risk component, but there is still room for hybridization. The "tails" of the distribution correspond to cases where two or more labels receive similar values $p_{d,l}$ – it is precisely these entries that can be interpreted as mixed incidents requiring special attention from analysts.

Next, the consistency of each semantic tag is calculated. At the same time, the size of each cluster is estimated $|C_{k,j}|$. The figure below shows the relationship between cluster sizes and their average semantic consistency. It can be seen that the resulting clusters vary greatly in size, which is natural for real operational data. Some incidents are routine events and occur frequently, while others are rare emergencies. Nevertheless, it is important to emphasize that the semantic consistency of the clusters remains at a sufficiently high level for all groups. This indicates that the selected taxonomy reflects objectively similar incident patterns and is not a random division. In other words, even small classes have semantic unity, which is extremely important for practical interpretation – such clusters correspond to specific but stable risk factors.

The resulting taxonomy of 15 classes can be interpreted as a hierarchy of risk incident categories. For clarity, here is a brief list of the semantic classes identified (the names assigned are based on an analysis of the top terms in each cluster):

- Inoperable System Performance Issues;
- Part 21 Reporting Issues;
- Generator Trip and Isolation;
- Plant Systems Maintenance & Response;
- Nuclear Regulatory Notifications;
- Nuclear Event Notifications;
- Fire Incident Notification;
- Hurricane-Related Plant Events;
- Seismic Event Response;
- Unusual Facility Atmosphere Events;
- Y-90 Treatment Dose Issues;
- Radiography Equipment Issues;

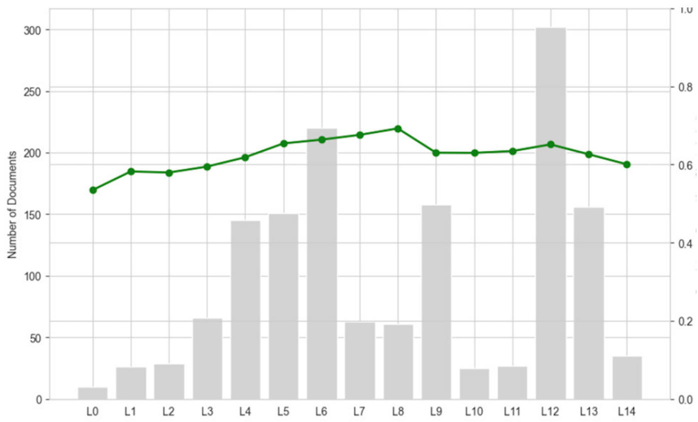

Fig. 3. Cluster sizes and their semantic consistency

- Lost Radioactive Gauge Reports;
- Radioactive Shipment Contamination;
- Missing or Lost Equipment.

This list reflects a cross-section of the latent structure of the corpus obtained by self-organization. It is crucial that the categories were not predetermined but emerged from the data. For example, the Hurricane-Related Events cluster grouped all messages related to hurricanes, even though they were not marked with a special tag in the original reports. Similarly, separate classes were formed around specific procedures (Part 21) or problems (loss of sensors), demonstrating the method's ability to detect implicit thematic associations in log entries.

The proposed method does not assume the existence of “correct” markup, so traditional metrics of accuracy and completeness are not directly applicable. Instead, the quality and usefulness of the resulting classification were evaluated according to three criteria: confidence distribution structure ξ_d, semantic consistency of clusters and topological isolation of clusters in feature space. The latter criterion quantitatively characterizes how geometrically separated clusters are and is calculated using the silhouette coefficient S:

$$S = \frac{b - a}{\max(a, b)},$$

where a is the average distance (1 – cosine proximity) between documents within a single cluster, b is the the smallest average distance from the documents in a given cluster to the documents in a neighboring cluster. In our experiment, we obtained the value $S \approx 0.02$, that is, formally $S > 0$. The small positive value of the silhouette is explained by the presence of many hybrid messages that create “bridges” between clusters and reduce intercluster distances. Nevertheless, $S > 0$ means non-zero isolation of the structure – the taxonomy has a certain stability despite the complexity of the data. It is important to note that discrete classification with this type of data would be methodologically incorrect. Many documents lie on the boundaries of classes and contain characteristics of several groups at the same time. That is why soft classification here is not a side effect, but a necessary way of describing reality. It recognizes the multi-causal nature of incidents and reflects it quantitatively.

It has been experimentally confirmed that the proposed method successfully solves the task of classifying risk incidents without a teacher. An interpretable taxonomy of 15 semantic categories is formed, and the classification confidence has a stable structure. The semantic consistency of clusters is maintained at a practical level of ~0.6, the cluster separation criterion S is positive, which indicates non-zero structural stability of the obtained taxonomy even with a significant proportion of hybrid events.

Discussion

The results obtained demonstrate the high effectiveness of the self-organizing risk incident classification method for analyzing large arrays of operational data. In traditional safety monitoring systems, each event is recorded under a single category, such as “operator error”, “equipment failure”, or “natural impact”. This simplification can lead to an underestimation of complex situations where several factors act together [22]. The proposed method allows, for the first time, an incident to be described by a set of risk factors with a quantitative assessment of each factor's contribution. The hybrid characterization of an incident provides more complete information for subsequent risk analysis. It is also consistent with the principles of Learning from Incidents (LFI) methodologies, which emphasize the need to consider the combination of causes and conditions of accidents [23]. Our method automates the multifactorial analysis of each case, which previously could only be achieved through labor-intensive expert analysis.

In practical deployment, the membership vector should be used as an operational risk profile for each incoming report. The largest component identifies the dominant factor, while the next components show secondary drivers that may escalate the event. A warning can be generated when the dominant score is at least 0.60 or when the sum of the two largest scores is at least 0.80, because this pattern indicates either a strong single threat or a stable hybrid case. Analysts can rank reports by these scores to prioritize response actions and investigation depth. Hybrid profiles are especially useful for root cause analysis, since they reveal interacting mechanisms such as equipment issues combined with procedural deviation. The same profiles support preventive planning, because recurring factor combinations can be tracked over time and mapped to targeted mitigation measures.

Unlike the “black boxes” of deep neural network models, the cluster-thematic taxonomy obtained by our method is interpretable by humans. Each class corresponds to an understandable semantic label (name), and the coefficients $p_{d,l}$ can be interpreted as the degree of influence of the relevant factors on the event d. This opens up the possibility of model validation by security experts. Moreover, the matrix $F = f_{d,l}$ can be considered as a normalized factor profile of incidents. Essentially, the method solves the problem of automatically forming a knowledge base about latent risk factors and their manifestations in real events. This result is directly relevant for proactive risk management – it allows you to build “threat maps” that show which combinations of factors occur most often and how they are interrelated. Similar ideas are embedded in ontological risk analysis systems, but there the factors are set manually. Here, they are extracted from the data, which confirms the thesis about the possibility of self-organization of the semantic structure of threats without a teacher.

Semantic class labels should be formed from the most representative terms of each leaf cluster and then embedded in the same normalized vector space as documents. A robust label can be built from top weighted terms that pass frequency and exclusivity filters, followed by expert screening to remove ambiguous tokens. To keep names stable after corpus updates, the previous label vector should be retained as an anchor and updated only when cosine similarity between old and new prototypes drops below a fixed threshold, for example 0.85. This rule prevents unnecessary renaming and preserves continuity of monitoring dashboards. The quality of label representation can be checked by measuring the gap between in cluster and out of cluster similarity to the label vector. Better label vectors improve the membership matrix because they increase contrast between dominant and secondary factors without forcing hard assignment.

Conclusion

The paper solves the problem of classifying risk incidents in the absence of prior labeling by means of self-organization of semantic clusters and soft classification of events according to them. A method for constructing a taxonomy of risk events directly from a text corpus has been developed. The taxonomy is a hierarchical structure of classes (topics and subtopics) identified without a teacher based on latent semantic analysis of data.

An approach to fuzzy classification of incidents is proposed, in which each message is assigned a vector of degrees of membership to all classes of the taxonomy. It is shown that such a continuous profile is resistant to minor text distortions and adequately reflects cases of hybrid incidents.

The interpretability and practical significance of the method has been experimentally confirmed using 10000 reports of events at nuclear power plants. An interpretable taxonomy of 15 categories has been constructed, which is consistent with expert opinions and complements them by identifying new combinations of factors. It has been shown that soft classification allows complex incidents to be broken down into several risk factors, which expands the information base for proactive risk management. This approach surpasses traditional single-class classification in terms of informativeness and flexibility.

Quality analysis was performed: semantic clusters have an average consistency of ~0.6, and the silhouette coefficient $S > 0$ confirms the preservation of the topological structure of clusters even with multiple intersections. This means that the selected classes are not random and can serve as a reliable basis for further monitoring and forecasting.

The method is fully automated and does not require labor-intensive data preparation. This is important for practical implementation in security monitoring systems, where the volume of unstructured data is constantly growing. The proposed algorithm can be integrated into problem-oriented decision support systems for filtering incident streams, early identification of dangerous trends, and developing risk management recommendations [24].

REFERENCES

1. **Makhutov N.A., Zatsarinny V.V., Algin V.B., Ishin N.N.** Technogenic risk, reliability and diagnostics of technical systems: approaches, models, methods. *Mechanics of Machines, Mechanisms and Materials*, 2012, Vol. 20–21, No. 3–4, Pp. 67–85.

2. **Riabinin I.A., Cherkesov G.N.** *Logiko-veroiatnostnye metody issledovaniia nadezhnosti strukturno-slozhnykh system* [*Logical and probabilistic methods for studying the reliability of structurally complex systems*]. Moscow: Radio i sviaz', 1981.

3. **Shannon C.E.** A Mathematical Theory of Communication. *The Bell System Technical Journal*, 1948, Vol. 27, pp. 379–423, 623–656.

4. **Wiener N.** *Cybernetics – Or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press, 1961.

5. **Zipf G.K.** *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley Press, 1949.

6. **Zhang J., Zi L., Hou Y., Deng D., Jiang W., Wang M.** A C-BiLSTM Approach to Classify Construction Accident Reports. *Applied Sciences*, 2020, Vol. 10, No. 17, Art. no. 5754. DOI: 10.3390/app10175754

7. **Chen Z., Huang K., Wu L., Zhong Z., Jiao Z.** Relational Graph Convolutional Network for Text-Mining-Based Accident Causal Classification. *Applied Sciences*, 2022, Vol. 12, No. 5, Art. no. 2482. DOI: 10.3390/app12052482

8. **Zeng W., Tang W., Yuan D., Zhang H., Duan P., Hu S.** Structure-Aware and Format-Enhanced Transformer for Accident Report Modeling. *Applied Sciences*, 2025, Vol. 15, No. 14, Art. no. 7928. DOI: 10.3390/app15147928

9. **Kim J.-M., Lim K.-K., Yum S.-G., Son S.** A Deep Learning Model Development to Predict Safety Accidents for Sustainable Construction: A Case Study of Fall Accidents in South Korea. *Sustainability*, 2022, Vol. 14, No. 3, Art. no. 1583. DOI: 10.3390/su14031583

10. **Xu H., Liu Y., Shu C.-M., Bai M., Motalifu M., He Z., Wu S., Zhou P., Li B.** Cause analysis of hot work accidents based on text mining and deep learning. *Journal of Loss Prevention in the Process Industries*, 2022, Vol. 76, Art. no. 104747. DOI: 10.1016/j.jlp.2022.104747

11. **Li J., Wu C.** Deep Learning and Text Mining: Classifying and Extracting Key Information from Construction Accident Narratives. *Applied Sciences*, 2023, Vol. 13, No. 19, Art. no. 10599. DOI: 10.3390/app131910599

12. **Qiang J., Qian Z., Li Y., Yuan Y., Wu X.** Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 2022, Vol. 34, No. 2, Pp. 1427–1445. DOI: 10.1109/TKDE.2020.2992485

13. **Wang Y., Zou P.X.W.** Decoding Construction Accident Causality: A Decade of Textual Reports Analyzed. *Buildings*, 2025, Vol. 15, No. 21, Art. no. 3859. DOI: 10.3390/buildings15213859

14. **Cao Y., Qu Z., Wu S., Chen Y., Skitmore M., Ma X., Wang J.** Analyzing OSHA Construction Accident Reports Using BERTopic Topic Modeling for Thematic Insights. *Buildings*, 2026, Vol. 16, No. 1, Art. no. 10. DOI: 10.3390/buildings16010010

15. **Markarian A.O., Meshcheryakov D.A., Gordeev G.Ya.** Structure and algorithm of the semantic analysis system for text data in the investigation of accidents and incidents at industrial facilities. *Legal Informatics*, 2025, Vol. 3, Pp. 57–68. DOI: 10.24412/1994-1404-2025-3-00-06

16. **Shrestha S., Morshed S.A., Pradhananga N., Lv X.** Leveraging Accident Investigation Reports as Leading Indicators of Construction Safety Using Text Classification. *ASCE Construction Research Congress* (*CRC*), 2020, Pp. 490–498. DOI: 10.1061/9780784482872.053

17. **Andrade S.R., Walsh H.S.** Machine learning framework for Hazard Extraction and Analysis of Trends (HEAT) in wildfire response. *Safety Science*, 2023, Vol. 167, Art. no. 106252. DOI: 10.1016/j.ssci.2023.106252

18. **Erfani A., Khanjar H.** Large Language Models for Construction Risk Classification: A Comparative Study. *Buildings*, 2025, Vol. 15, No. 18, Art. no. 3379. DOI: 10.3390/buildings15183379

19. **Zhang L., Hou Y., Ren F.** AIR Agent: A GPT-Based Subway Construction Accident Investigation Report Analysis Chatbot. *Buildings*, 2025, Vol. 15, No. 4, Art. no. 527. DOI: 10.3390/buildings15040527

20. **Du G., Chen A.** Coal Mine Accident Risk Analysis with Large Language Models and Bayesian Networks. *Sustainability*, 2025, Vol. 17, No. 5, Art. No. 1896. DOI: 10.3390/su17051896

21. **Rodionov D.G., Polyakov P.A., Yakob P.A., Konnikov E.A.** A scenario-optimisation method for semantic processing of textual data on accidental incidents at industrial enterprises. *Soft Measurement and Computing*, 2025, Vol. 95, No. 10, Pp. 117–129. DOI: 10.36871/2618-9976.2025.10.009

22. **Woźniak Z., Hoła B.** Analysing Near-Miss Incidents in Construction: A Systematic Literature Review. *Applied Sciences*, 2024, Vol. 14, No. 16, Art. no. 7260. DOI: 10.3390/app14167260

23. **Guan J., Xu Y., Chan A.P.C., Choi T., Yang Y.** Factors affecting learning from incidents: A cross-industry review. *Journal of Loss Prevention in the Process Industries*, 2024, Vol. 89, Art. no. 105297. DOI: 10.1016/j.jlp.2024.105297

24. **Suh Y.** Identifying Safety Technology Opportunities to Mitigate Safety-Related Issues on Construction Sites. *Buildings*, 2025, Vol. 15, No. 6, Art. no. 847. DOI: 10.3390/buildings15060847

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Viacheslav P. Shkodyrev
Шкодырев Вячеслав Петрович
E-mail: shkodyrev@mail.ru

Dmitry G. Rodionov
Родионов Дмитрий Григорьевич
E-mail: drodionov@spbstu.ru
ORCID: https://orcid.org/0000-0002-1254-0464

Evgenii A. Konnikov
Конников Евгений Александрович
E-mail: konnikov.evgeniy@gmail.com
ORCID: https://orcid.org/0000-0002-4685-8569

Submitted: 26.12.2025; Approved: 13.03.2026; Accepted: 17.03.2026.
Поступила: 26.12.2025; Одобрена: 13.03.2026; Принята: 17.03.2026.