Contents

Semantic-Enriched Mining Accident Analysis Using SBERT Embeddings, UMAP Dimensionality Reduction, and K-Means Clustering for Pattern Discovery

aakdani1@gmail.com
aakdani1@gmail.com

June 21, 2025

Abstract

This study presents an integrated framework for mining accident pattern discovery in the mining sector by combining advanced natural language processing techniques with structured metadata analysis. Utilizing pre-trained Sentence-BERT (SBERT) models, dense and semantically rich sentence embeddings are generated from unstructured accident narratives, capturing nuanced contextual information. These embeddings are enriched with metadata variables from the Mine Safety and Health Administration (MSHA) database to form comprehensive feature representations. To address the challenges posed by high-dimensional data, Uniform Manifold Approximation and Projection (UMAP) is employed for dimensionality reduction, preserving both local and global semantic structures. Subsequently, k-means clustering is applied to the reduced embeddings to identify latent patterns and groupings within the accident data. The approach enhances the interpretability and accuracy of accident causation analysis beyond traditional methods such as TF-IDF and word-level embeddings. Post-clustering analyses reveal meaningful correlations between narrative themes and metadata attributes, supporting targeted risk management and prevention strategies. The framework demonstrates scalability, computational efficiency, and adaptability, with potential applications across various safety-critical industries. Future directions include multimodal data integration, improved interpretability, and real-time monitoring capabilities to advance occupational safety research and practice.

1 Introduction

The analysis of accident narratives in mining has evolved significantly with the advent of advanced natural language processing (NLP) techniques and machine learning models. Traditional approaches, such as content analysis and manual coding, have provided valuable insights into accident causation and risk patterns, but they are often limited by scalability and the ability to capture nuanced semantic relationships within unstructured text [1][2]. The integration of computational methods, including the use of software and models, has become increasingly prevalent in recent years, with a notable proportion of studies leveraging these tools to enhance the understanding and prediction of mining accidents [1]. Recent developments in NLP have introduced sophisticated embedding models capable of generating dense, high-dimensional vector representations of textual data. These embeddings encapsulate rich semantic information, enabling more precise analysis of accident reports and facilitating the identification of latent patterns that may not be apparent through traditional methods [3][4]. The use of pre-trained models, such as SBERT, allows for the extraction of sentence-level embeddings that are contextually informed, capturing both syntactic and semantic nuances present in accident narratives [5]. By combining these embeddings with structured metadata from sources like the MSHA accident database, researchers can construct comprehensive feature sets that reflect both the textual and contextual dimensions of each incident [3]. However, the high dimensionality of these embeddings poses challenges for direct analysis and visualization. Dimensionality reduction techniques, such as Uniform Manifold Approximation and Projection (UMAP), have emerged as effective solutions for transforming complex, high-dimensional data into lower-dimensional spaces while preserving the intrinsic structure and relationships among data points. UMAP has demonstrated superior performance compared to other reduction methods like Principal Component Analysis (PCA) and t-Distributed

Stochastic Neighbor Embedding (t-SNE), particularly in its ability to generate clear and interpretable visualizations of clusters, even with large and complex datasets. The stability and reliability of UMAP across various reduction ratios further enhance its suitability for mining accident data, where the underlying patterns may be subtle and multifaceted. Once the embeddings are projected into a lower-dimensional space, clustering algorithms such as k-means can be applied to group similar incidents based on their semantic and contextual features [6]. This clustering facilitates the discovery of patterns and trends within the accident data, supporting a more granular examination of causative factors and risk phenomena [2][3]. The combination of semantic-rich embeddings, robust dimensionality reduction, and clustering analysis represents a methodological advancement that extends beyond the capabilities of traditional statistical or rule-based approaches [2]. For instance, deep learning models tailored to the construction and mining industries have shown improved accuracy in classifying and interpreting accident causes, offering deeper insights into the mechanisms underlying these events [4]. The integration of these techniques not only enhances the analytical depth but also supports the development of frameworks that can inform policy, safety management, and preventive strategies in mining operations [3][7]. By leveraging the strengths of NLP, dimensionality reduction, and clustering, researchers are equipped to extract actionable knowledge from vast repositories of accident narratives, ultimately contributing to improved safety outcomes and risk mitigation in the mining sector [1][7].

2 Background and Motivation

2.1 Challenges in Accident Data Analysis

Accident data analysis is confronted with a multitude of challenges, many of which stem from the inherent complexity and unstructured nature of the data itself. Accident reports, particularly in the mining and construction sectors, are often composed of free-text narratives that lack standardized formats, making automated extraction of meaningful information a non-trivial task [8][1]. The absence of structured safety risk data from these descriptive contents leads to a reliance on subjective judgments, which in turn diminishes the reliability and reproducibility of research outcomes. Furthermore, the relationships among various accident participants, as well as the distinction between direct and indirect risk factors, are frequently ambiguous or insufficiently documented, complicating efforts to establish causal links and comprehensive risk models [8][9]. The high dimensionality of textual data, especially when processed using modern embedding techniques such as SBERT, introduces additional analytical hurdles. Embedding models generate dense, high-dimensional vectors, often with thousands of features per narrative, which are difficult to interpret and computationally expensive to analyze directly [2]. This phenomenon, commonly referred to as the curse of dimensionality, adversely affects the performance of traditional clustering algorithms, such as k-means, and density-based methods, which struggle to discern meaningful groupings in such spaces [10][2]. Radu et al. [10] indicate that density-based algorithms are particularly sensitive to high dimensionality, but their performance can be improved when dimensionality is reduced and semantic-rich representations like Doc2Vec are employed. Dimensionality reduction techniques, such as UMAP, t-SNE, and PCA, have been introduced to address these issues by projecting high-dimensional embeddings into lower-dimensional spaces while preserving essential structural relationships [6][11]. However, each method presents its own set of challenges. For instance, PCA may distribute data points in a manner that obscures cluster boundaries, whereas t-SNE and UMAP are more effective at revealing latent clusters but can be sensitive to parameter choices and may not always preserve global data structure [6][11]. Hozumi et al. [6] demonstrate that dimension-reduced k-means clustering not only improves computational efficiency but also enhances clustering accuracy, yet the selection of appropriate reduction ratios and the interpretability of resulting clusters remain open questions. Another significant challenge lies in the pre-processing of accident narratives. The process of lemmatization, removal of stop words, and normalization of terminology is essential to reduce noise and redundancy in the data, but it can inadvertently strip away context or introduce artifacts, as seen in the transformation of words like "install" to "instal". The sheer diversity of vocabulary, thousands of unique words across accident categories, further complicates the construction of robust feature spaces for analysis [12]. Moreover, the integration of metadata variables, such as accident type, location, and time, with semantic embeddings requires careful alignment to ensure that both structured and unstructured information contribute meaningfully to downstream analyses [2][3]. The interpretability of clustering results poses yet another obstacle. While advanced clustering of embeddings can reveal patterns and trends in accident causes, translating these clusters into actionable insights for safety management and policy development is not straightforward [2][3][1]. The lack of clear, observable precursors to injuries, as highlighted by Tixier et al. [9], underscores the difficulty in distinguishing between antecedent conditions and outcomes within narrative data. This ambiguity can lead to misclassification or oversimplification of complex accident scenarios. Despite the proliferation of software tools and analytical models, a substantial proportion of studies still rely on manual or semi-automated content analysis, which is labor-intensive and susceptible to human bias [1]. The integration of advanced natural language processing (NLP) techniques, such as word embeddings and topic modeling, has shown promise in automating the extraction and classification of safety-relevant information, yet these approaches are not without limitations. For example, the effectiveness of clustering algorithms is contingent upon the quality of the underlying embeddings and the appropriateness of the chosen similarity metrics, such as Euclidean distance in high-dimensional spaces [2]. In summary, accident data analysis is challenged by the unstructured nature of narrative reports, high dimensionality of semantic representations, difficulties in pre-processing and feature integration, and the interpretability of clustering outcomes. Addressing these challenges requires a combination of advanced NLP techniques, robust dimensionality reduction, and careful methodological design to ensure that the resulting insights are both reliable and actionable [6][10][2][8][9][3][1][12][11].

2.2 Significance of Mining Accident Reports

Mining accident reports are a critical resource for understanding the underlying causes, mechanisms, and consequences of incidents within the mining sector. These reports, often comprising detailed narratives and structured metadata, serve as a foundation for both qualitative and quantitative analyses aimed at improving occupational safety and operational efficiency. By systematically analyzing accident scenarios, researchers can uncover patterns and trends that may not be immediately apparent, thereby enriching the comprehension of safety risks associated with mining activities [1]. The integration of computational methods, such as text mining and machine learning, has further enhanced the ability to extract meaningful information from these reports, enabling automated classification and identification of contributory factors [13]. The significance of mining accident reports extends beyond mere documentation; they are instrumental in informing risk assessment, regulatory compliance, and the development of targeted intervention strategies. For instance, content analysis of these reports can reveal previously unnoticed correlations between specific job groups, equipment types, and accident outcomes, which is essential for designing effective preventive measures [14]. The use of advanced software and analytical models has become increasingly prevalent, with a substantial proportion of studies leveraging these tools to analyze accident causes and predict future incidents [1]. This shift towards data-driven approaches underscores the value of mining accident reports as a rich source of empirical evidence for safety research. Text mining techniques applied to accident narratives facilitate the identification of significant terms and phrases, laying the groundwork for computational frameworks that automate the categorization of reports based on contributing factors [13]. Such automation not only enhances the efficiency and consistency of accident analysis but also supports the creation of taxonomies that can be integrated with established systems like the ICAO Accident/Incident Data Reporting (ADREP) taxonomy, albeit with varying levels of granularity [15]. The ability to classify accidents and extract semantic relationships from narrative data represents a significant advancement over traditional manual review processes [4][16]. Furthermore, mining accident reports provide valuable insights into the operational context and environmental conditions surrounding incidents. For example, analyses have shown that maintenance personnel and employees operating mining machines are at higher risk of accidents resulting in significant lost workdays, particularly in surface plants and coal handling facilities [14]. Such findings highlight the importance of targeted safety interventions and resource allocation. The application of clustering and dimensionality reduction techniques to high-dimensional embeddings derived from accident narratives enables the discovery of latent structures and groupings within the data [6]. This approach allows for a more nuanced understanding of accident causality and the identification of emergent trends that may inform future safety protocols. The combination of semantic-rich embeddings with metadata variables further enhances the granularity of analysis, providing a comprehensive view of the factors contributing to mining accidents [2]. In summary, mining accident reports are indispensable for advancing the science of accident analysis and prevention. Their systematic examination, supported by modern computational techniques, not only deepens the understanding of accident causation but also drives the development of more effective safety management systems [1][13][16].

2.3 Limitations of Traditional Analytical Techniques

Traditional analytical techniques for mining accident data, such as frequency analysis and basic statistical modeling, exhibit several notable limitations when applied to complex, high-dimensional narrative datasets. One of the primary challenges is their inability to capture the nuanced semantic relationships embedded within textual accident reports. Frequency-based approaches, for instance, tend to emphasize the most commonly reported events, which can obscure less frequent but potentially critical patterns. This is particularly problematic in domains like mining safety, where rare but severe incidents may be overshadowed by the sheer volume of routine reports, leading to an incomplete understanding of underlying risk factors [15]. Dimensionality reduction and clustering are essential for managing and interpreting high-dimensional data, yet traditional methods such as Principal Component Analysis (PCA) and classical multidimensional scaling (MDS) are often inadequate for datasets characterized by large vector magnitudes and complex semantic structures. These techniques typically assume linear relationships and may fail to preserve the intricate, non-linear dependencies present in narrative data, resulting in loss of important contextual information [2]. Furthermore, conventional word embedding methods, which rely on pre-defined vocabularies, struggle to handle out-of-vocabulary or domain-specific terms frequently encountered in mining accident narratives. This limitation restricts their applicability to word-level tasks and hinders their effectiveness in sentence-level analyses, such as sentiment detection or classification of accident types [3]. The reliance on absolute counts and simple correlations in traditional analyses also limits the discovery of meaningful metadata patterns. For example, when pilot reports are analyzed solely by frequency, the overwhelming number of common events can mask the presence of sub-clusters related to specific operational issues, such as maintenance or ground traffic. Only through more sophisticated correlation-based studies can these subtle patterns be revealed, highlighting the inadequacy of traditional methods for uncovering deeper insights [15]. Additionally, classical models often fail to integrate heterogeneous data sources, such as combining narrative text with structured metadata, which is crucial for comprehensive accident analysis. Another significant drawback is the computational inefficiency and scalability issues associated with some traditional models. As datasets grow in size and complexity, these methods become increasingly impractical, both in terms of processing time and memory requirements. For instance, while some sentence embedding models offer improved semantic representation, their computational speed on standard hardware can be a bottleneck, especially when compared to more advanced architectures designed for efficiency. Reimers et al. [17] highlight that even among neural models, architectural choices can lead to substantial differences in processing speed, further complicating the use of traditional techniques for large-scale analyses. Moreover, traditional analytical frameworks often lack mechanisms to capture long-range dependencies and contextual relationships within narratives. Models based solely on local context, such as those using only short-term word co-occurrences, are insufficient for understanding the broader semantic structure of accident reports. Advanced architectures that incorporate self-attention mechanisms or recurrent neural networks, such as BiLSTM or BiGRU, have demonstrated superior ability to model these dependencies, underscoring the limitations of earlier approaches [18][19]. The interpretability and practical utility of results derived from traditional techniques are also limited. Visual representation methods, while helpful in reducing analyst workload, may not fully exploit the semantic richness of narrative data unless combined with advanced feature extraction and dimensionality reduction strategies [20]. Furthermore, the inability to effectively handle unregistered or domain-specific vocabulary restricts the generalizability of traditional word embedding approaches, making them less suitable for specialized domains like mining safety [3]. In summary, the constraints of traditional analytical techniques, ranging from their reliance on frequency-based metrics and linear dimensionality reduction, to their limited handling of semantic complexity and computational inefficiency, necessitate the adoption of more sophisticated frameworks. Integrating semantic-rich embeddings, advanced dimensionality reduction, and clustering methods offers a pathway to overcome these challenges and achieve deeper, more actionable insights into mining accident patterns [15][2][17][18][20][19][3].

3 Review of Text Mining in Accident Analysis

3.1 Overview of Natural Language Processing in Safety Research

Natural Language Processing (NLP) has become a cornerstone in the analysis of safety-related textual data, particularly in the context of accident reports and narratives. The application of NLP in safety research is driven by the need to extract actionable knowledge from unstructured text, which is abundant in accident documentation across various industries, including mining, construction, and chemical processing [2][16][21]. Accident narratives, which are often detailed and context-rich, provide a valuable source of information for understanding the underlying causes and contributing factors of incidents. The authors of [2] indicate that the examination of these narratives using NLP techniques such as text classification and mining enables researchers to systematically extract and analyze critical information that would otherwise remain hidden in free-text fields. Text mining, a subset of NLP, is particularly effective in identifying patterns, trends, and relationships within large corpora of accident reports. By leveraging advanced algorithms, researchers can uncover previously unnoticed risk factors and emergent themes that inform safety management strategies [1][16]. For instance, content analysis supported by software tools allows for the efficient processing of vast datasets, facilitating the discovery of trends that may not be apparent through manual review alone [1]. This computational approach not only accelerates the analysis process but also enhances the reproducibility and objectivity of findings. The integration of NLP with machine learning models has further advanced the field. Techniques such as document embedding, where textual data is transformed into high-dimensional vectors, enable the application of clustering and classification algorithms to group similar incidents and identify commonalities [10][2]. Radu et al. [10] state that models like Doc2Vec improve the compactness of clusters and preserve the relationship between documents and their topics, which is crucial for meaningful categorization of accident reports. Smetana et al. [2] outline that clustering embeddings based on their similarities allows for a nuanced examination of accident causes, with dense vector representations supporting sophisticated analytical methods. Recent developments in deep learning, particularly the advent of large language models (LLMs), have significantly enhanced the capabilities of NLP in safety research. LLMs, characterized by their vast parameter space and extensive training on diverse text corpora, demonstrate remarkable performance in tasks such as text classification, summarization, and semantic analysis [4]. These models can process and understand complex language structures, making them well-suited for extracting insights from accident narratives that may contain domain-specific terminology and nuanced descriptions of events. The utility of NLP extends beyond mere classification or clustering. It enables the identification of high-frequency accident types, common causes, and precursors, which are essential for risk analysis and the development of targeted safety interventions [22]. By systematically analyzing accident narratives, safety professionals can gain a comprehensive understanding of the factors contributing to incidents, thereby informing the design of more effective prevention strategies [21]. Moreover, the application of NLP in mining accident analysis, though less common compared to other industries, has shown promise in extracting meaningful patterns from fatal accident narratives, such as identifying frequently mentioned objects or conditions associated with accidents [16]. The flexibility and adaptability of NLP methods are further exemplified by their ability to accommodate language-specific characteristics and integrate with ontological frameworks, as suggested for future research directions [8]. This adaptability ensures that NLP techniques remain relevant and effective across different linguistic and regulatory contexts. In summary, the adoption of NLP in safety research represents a significant methodological advancement, enabling the systematic extraction of knowledge from unstructured accident data. The combination of semantic-rich embeddings, dimensionality reduction, and clustering provides a robust framework for uncovering deep insights into accident causation and prevention, surpassing the limitations of traditional manual analysis [2][16]. The ongoing evolution of NLP, driven by advances in machine learning and deep learning, continues to expand its potential for enhancing safety outcomes across diverse industrial domains.

3.2 Text Representation Methods

3.2.1 Bag-of-Words and TF-IDF

Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are foundational text representation techniques that have been widely adopted in accident analysis studies. The BoW approach transforms textual narratives into a structured format by representing each document as a vector of word counts, disregarding grammar and word order but preserving multiplicity. This method enables the conversion of unstructured accident reports into a format amenable to statistical and machine learning analyses, facilitating the identification of frequently occurring terms and their distribution across the dataset. By reducing narratives to a matrix of word frequencies, BoW provides a straightforward yet effective means to quantify textual data, which is particularly useful for initial exploratory analysis and clustering tasks. TF-IDF extends the BoW model by incorporating the relative importance of words within and across documents. While BoW captures raw frequency, TF-IDF adjusts these counts by penalizing common words and amplifying the significance of terms that are distinctive to specific documents. This weighting mechanism is crucial in accident analysis, as it helps to highlight unique factors or rare events that may be critical for understanding causation or identifying emerging trends. The TF-IDF matrix, therefore, serves as a more discriminative representation, allowing for improved differentiation between accident narratives that may otherwise appear similar in a simple frequency-based model. The practical application of BoW and TF-IDF in accident analysis often involves several preprocessing steps, such as tokenization, lower-casing, stop-word removal, and stemming, to standardize the text and reduce dimensionality. These steps ensure that the resulting matrices are both computationally manageable and semantically meaningful. Once the BoW or TF-IDF matrices are constructed, they can be subjected to further analysis, such as k-means clustering, to group similar accident reports and uncover latent patterns within the data. Visualization techniques like t-SNE are sometimes employed to project these high-dimensional representations into lower-dimensional spaces, aiding in the interpretation of clustering results and the identification of key trends. Despite their simplicity, BoW and TF-IDF remain valuable tools in the text mining toolkit for accident analysis. They provide a transparent and interpretable means of representing narrative data, which is essential for both exploratory data analysis and the development of more advanced models. However, these methods are limited in their ability to capture semantic relationships and contextual nuances, motivating the integration of more sophisticated embedding techniques in recent research. Nevertheless, the foundational role of BoW and TF-IDF in structuring and quantifying accident narratives continues to underpin many analytical workflows in the field [15].

3.2.2 Word Embedding Approaches

Word embedding approaches have become fundamental in the representation of textual data for accident analysis, enabling the transformation of unstructured narratives into structured, high-dimensional vectors that capture semantic and syntactic relationships. The core idea behind word embeddings is to map words, sentences, or even entire documents into continuous vector spaces, where the proximity of vectors reflects semantic similarity. This mapping facilitates computational manipulation and analysis of text, which is essential for extracting meaningful patterns from large-scale accident datasets [3]. Traditional word embedding models, such as Word2Vec and GloVe, generate dense vector representations by leveraging the co-occurrence statistics of words within a corpus. These models are capable of capturing both syntactic and semantic relationships, allowing for the quantification of word similarity and the identification of latent structures in text data [2]. The use of pre-trained embeddings, which are trained on large external corpora, has been shown to enhance the performance of downstream tasks by providing rich contextual information that may not be present in smaller, domain-specific datasets. For instance, Fan Zhang et al. [11] demonstrate that seeding embedding layers with pre-trained word vectors, particularly when using bigram contexts and higher dimensionality settings, leads to improved classification performance as measured by weighted average F1 scores. Beyond word-level embeddings, sentence and document embeddings extend this concept to larger textual units, enabling the representation of entire accident narratives as single vectors. This is particularly advantageous in accident analysis, where the context and sequence of events are often critical for understanding causation and identifying trends [3]. Pre-trained models such as BERT and SBERT have gained prominence due to their ability to generate context-aware embeddings that reflect the meaning of sentences within their broader textual context. These models utilize deep neural architectures to encode not only the local context of words but also the global structure of sentences, resulting in embeddings that are well-suited for clustering, classification, and similarity analysis [2]. The dimensionality of embeddings is a crucial parameter that influences both the expressiveness and computational efficiency of the representation. Higher-dimensional embeddings can capture more nuanced relationships but may also introduce redundancy and increase the risk of overfitting, especially in smaller datasets [11]. Dimensionality reduction techniques, such as UMAP, are often employed to project high-dimensional embeddings into lower-dimensional spaces, preserving the essential topological structure while facilitating visualization and clustering. Yuta Hozumi et al. [6] outline that optimizing the spectral layout of data in a low-dimensional space minimizes the error between the original and projected topologies, which is particularly beneficial for large-scale analyses. In the context of accident analysis, embedding approaches have been integrated with clustering algorithms to uncover patterns and categorize incidents based on their semantic content. The process typically involves generating embeddings for each narrative, reducing their dimensionality, and then applying clustering methods such as k-means to group similar cases. This workflow enables the identification of major causes and trends in accident data, supporting both descriptive and predictive analytics [2][16]. Mason Smetana et al. [2] emphasize that the use of high-dimensional embeddings, such as those with 1536 dimensions, necessitates advanced analytical methods to manage the complexity and extract actionable insights. Moreover, the integration of metadata variables with semantic-rich embeddings further enhances the analytical framework. By combining structured information (e.g., accident type, location, time) with unstructured narrative embeddings, researchers can perform more comprehensive analyses that account for both contextual and categorical factors. This hybrid approach supports the identification of underlying relationships between textual and non-textual data, offering a more holistic understanding of accident causation and trends [15][1]. Automated feature extraction methods, including convolutional neural networks (CNNs), have also been explored as alternatives to manual or traditional feature selection techniques such as TF-IDF. These methods can efficiently extract salient features from text without requiring expert intervention, streamlining the analysis pipeline and enabling scalability to larger datasets [20]. The combination of embedding-based representations with machine learning classifiers and clustering algorithms has been shown to outperform traditional approaches in various text mining tasks related to safety and accident analysis [2][20]. In summary, word embedding approaches provide a robust foundation for representing and analyzing accident narratives. By capturing semantic relationships and enabling integration with metadata, these methods facilitate advanced clustering and pattern discovery, supporting the development of innovative frameworks for accident analysis that surpass the capabilities of conventional text mining techniques [11][2][3].

3.3 Topic Modeling and Dimensionality Reduction

3.3.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a widely adopted probabilistic topic modeling technique that has been extensively utilized for extracting latent thematic structures from large collections of textual data, including accident narratives. LDA conceptualizes each document as a mixture of topics, where each topic is characterized by a distribution over words. This generative process enables the model to uncover hidden semantic patterns within the corpus, facilitating the identification of recurring themes that may not be immediately apparent through manual inspection. The operational mechanism of LDA involves a three-phase stochastic approach, where topics are allocated to clusters of words within each document, thereby capturing the underlying thematic structure of the narratives [23]. The application of LDA in accident analysis has demonstrated its utility in revealing the latent topics embedded within unstructured text data, such as safety reports and incident descriptions. For instance, LDA has been employed to analyze aviation safety reports, where it transformed the document-term matrix into a lower-dimensional space, effectively uncovering relationships between terms and documents. This dimensionality reduction not only aids in topic extraction but also provides insights into the latent themes present in the corpus, which can be instrumental for safety management and risk assessment [13]. The ability of LDA to model documents as distributions over topics allows for a nuanced understanding of the multifaceted nature of accident causation and reporting. Comparative studies have highlighted the complementary nature of LDA and other topic modeling techniques, such as Latent Semantic Analysis (LSA), in the context of accident data classification. Williams and Betak, as referenced by Bridgelall et al., found that both LDA and LSA were effective in identifying critical issues, such as the prevalence of tractor-trailer trucks at rail grade crossings, within railroad accident reports. The integration of text mining and spatial analysis further enhanced the identification of high-risk locations, underscoring the practical value of LDA in real-world safety applications [24]. This suggests that LDA, when combined with other analytical methods, can contribute to a more comprehensive understanding of accident patterns and inform targeted intervention strategies. The performance of LDA is influenced by the choice of input representations, such as term frequency-inverse document frequency (TF-IDF) and distributed embeddings like Doc2Vec. Radu et al. report that LDA, when applied to TF-IDF and Doc2Vec representations, yields varying levels of clustering performance, as measured by standard evaluation metrics. Notably, embeddings generated by models such as Doc2Vec are independent of the preprocessing method, which may offer additional flexibility in integrating LDA with modern embedding techniques [10]. This adaptability is particularly relevant in contemporary frameworks that leverage pre-trained language models to generate semantically rich representations of accident narratives. The integration of LDA with dimensionality reduction techniques and clustering algorithms further enhances its analytical capabilities. For example, the combination of LDA with clustering methods enables the grouping of documents based on their topic distributions, facilitating the identification of patterns and trends across large-scale accident datasets. This approach aligns with the broader movement in accident analysis towards automated, data-driven methodologies that can efficiently process and interpret vast amounts of unstructured text [13][24][10][23]. The probabilistic graphical representation of LDA, as illustrated in recent studies, provides a transparent and interpretable framework for topic extraction, which is essential for communicating findings to stakeholders in safety-critical domains [13][23]. In summary, LDA serves as a foundational technique in the review of text mining approaches for accident analysis, offering robust mechanisms for topic discovery, dimensionality reduction, and thematic clustering. Its probabilistic modeling framework, compatibility with various input representations, and synergy with other analytical methods position LDA as a valuable tool for extracting actionable insights from complex accident narrative data [13][24][10][23].

3.3.2 Structural Topic Modeling (STM)

Structural Topic Modeling (STM) represents an advanced approach within the broader family of topic modeling techniques, designed to uncover latent thematic structures in large collections of textual data while explicitly incorporating document-level metadata. Unlike traditional models such as Latent Dirichlet Allocation (LDA) or Latent Semantic Analysis (LSA), STM enables the integration of external variables, such as time, location, or categorical attributes, directly into the generative process, allowing for the examination of how topics vary with respect to these covariates. This capacity is particularly relevant in accident analysis, where metadata from sources like the MSHA accident database can provide critical context for interpreting narrative content. The core mechanism of STM involves representing each document as a mixture of topics, where each topic is a distribution over words. However, STM extends this framework by modeling topic prevalence and content as functions of document-level covariates. This means that the probability of a topic appearing in a document, as well as the specific vocabulary used to express that topic, can systematically vary according to metadata such as accident type, mine size, or worker experience. Such flexibility is not present in standard LDA, which assumes that topic distributions are independent of external variables [13][20]. In the context of accident narrative analysis, STM offers several advantages. By leveraging both the semantic information captured in sentence embeddings and the structured metadata, STM can reveal nuanced patterns that may be obscured in models that ignore contextual variables. For example, STM can identify whether certain safety issues are more prevalent in specific types of mines or under particular operational conditions, and how the language used to describe these issues shifts across different contexts [1]. This is particularly valuable for mining accident data, where factors such as job change, lack of routine, or mine size have been shown to influence risk profiles [25]. The integration of STM with modern embedding techniques, such as those generated by pre-trained SBERT models, further enhances its analytical power. Embeddings provide dense, semantic-rich representations of accident narratives, capturing subtle relationships between words and phrases that traditional bag-of-words models may miss [3][2]. When these embeddings are combined with STM, the resulting framework can model both the latent thematic structure and the influence of metadata, offering a more comprehensive understanding of accident causation and trends. Dimensionality reduction techniques, such as UMAP, play a complementary role by projecting high-dimensional embeddings into lower-dimensional spaces suitable for clustering and visualization [6]. This step facilitates the identification of coherent topic clusters and the exploration of their relationships with metadata variables. For instance, UMAP can help reveal how accident narratives group according to underlying themes, while STM can explain how these themes are modulated by factors like shift length or contractor status [25][6]. The application of STM in accident analysis is further supported by the growing use of software tools and models in the field. As noted in recent reviews, a significant proportion of studies now employ computational methods to analyze accident causes, predict future incidents, and support prevention strategies [1]. STM, with its ability to incorporate both text and metadata, aligns well with this trend and addresses the limitations of purely text-based or metadata-driven approaches. Moreover, STM facilitates the quantification of topic co-occurrence and the construction of factor-accident matrices, which are essential for modeling the relationships between contributing factors and accident outcomes [3]. By capturing the joint distribution of topics and metadata, STM enables the identification of complex interaction patterns that may underlie accident causation, supporting more targeted interventions and policy recommendations. The authors of [13] indicate that dimensionality reduction techniques like LSA have been successfully used to extract latent themes from safety reports, providing insights into the underlying structure of accident data. STM builds upon these foundations by allowing for the explicit modeling of how these themes are influenced by document-level variables, thus offering a richer and more interpretable framework for topic discovery. In summary, STM represents a significant methodological advancement for text mining in accident analysis. By integrating semantic-rich embeddings, dimensionality reduction, and metadata-aware topic modeling, STM provides a robust framework for uncovering patterns and trends in accident narratives that are closely linked to contextual factors. This approach not only enhances the interpretability of topic models but also supports the development of more effective safety interventions and risk management strategies [3][1][13][25].

3.3.3 Emergence of Neural Embedding Models

The emergence of neural embedding models has significantly transformed the landscape of text mining, particularly in the context of accident analysis. Traditional approaches to representing textual data, such as bag-of-words or TF-IDF, often fail to capture the nuanced semantic relationships inherent in natural language. Neural embedding models, by contrast, encode words, sentences, or even entire documents into dense, continuous vector spaces, where semantic similarity is reflected in geometric proximity. This paradigm shift enables more sophisticated analyses of accident narratives, as the embeddings encapsulate both syntactic and semantic information [26][2]. Early neural embedding techniques, such as Word2Vec and GloVe, introduced the concept of distributed representations, where each word is mapped to a high-dimensional vector based on its context within large corpora. These models excel at capturing word-level relationships and have been widely adopted for various natural language processing (NLP) tasks. For instance, the skip-gram model of Word2Vec is recognized for its ability to generate powerful word embeddings that are particularly suitable for downstream applications, including clustering and classification. The resulting vector matrices serve as a foundation for more advanced feature extraction, often through convolutional neural networks, which can further distill high-level semantic concepts from the embedded representations [26]. However, word-level embeddings are limited in their capacity to represent the meaning of entire sentences or documents, especially when context and word order play crucial roles. To address these limitations, the field has witnessed the development of sentence and document embedding models. Pre-trained transformer-based architectures, such as BERT and RoBERTa, have been fine-tuned to produce sentence embeddings that capture richer contextual information. These models leverage deep bidirectional attention mechanisms, allowing them to encode complex dependencies and semantic nuances across sentences [17][2]. Reimers et al. [17] highlight that starting from pre-trained transformer models and fine-tuning them for sentence embedding tasks yields representations that are highly effective for semantic similarity and clustering, surpassing earlier methods that relied on random initialization. The integration of neural embedding models into accident analysis workflows enables the transformation of unstructured narrative data into structured, high-dimensional vectors. This process facilitates the application of advanced analytical techniques, such as clustering and dimensionality reduction, which are essential for uncovering latent patterns and trends in accident datasets. For example, embedding models can generate dense vectors of substantial dimensionality (e.g., 1536 dimensions), necessitating the use of sophisticated methods for subsequent analysis and visualization. The ability to cluster these embeddings based on their similarities allows for a more granular examination of the underlying causes and contributing factors in accident reports [2]. Moreover, the adoption of neural embedding models aligns with the broader trend of leveraging software and computational models to enhance the analysis of accident scenarios. The use of such models not only streamlines the extraction of meaningful patterns from large-scale textual data but also supports the development of predictive and preventive strategies in safety-critical domains [1]. By encoding accident narratives into semantically rich vectors, researchers can systematically explore the relationships between incidents, identify emerging hazards, and inform policy and operational decisions [3]. In summary, the rise of neural embedding models marks a substantial advancement in the field of text mining for accident analysis. These models provide a robust framework for representing and analyzing complex narrative data, enabling deeper insights into the mechanisms and trends underlying mining accidents [26][17][1][2].

4 Integration of Structured and Unstructured Data

4.1 Nature of Accident Narratives and Metadata

Accident narratives, as found in mining safety reports, are unstructured textual descriptions that encapsulate the sequence of events, contextual factors, and causal mechanisms underlying each incident. These narratives are rich in semantic content, often providing nuanced details that structured data fields may overlook. The examination of such narratives is essential for a comprehensive understanding of safety risks, as they contain information about the circumstances, actions, and environmental conditions that precede and follow an accident. By leveraging natural language processing (NLP) techniques, researchers can systematically extract and analyze this wealth of information, moving beyond the limitations of manual review and basic keyword searches [2]. The integration of accident narratives with structured metadata variables, such as those provided by the MSHA (Mine Safety and Health Administration) accident data, enables a more holistic analysis. Metadata typically includes categorical and numerical fields such as accident type, location, time, equipment involved, and worker demographics. These structured variables offer a standardized framework for comparison and statistical analysis, but they may lack the contextual richness present in narrative text [3]. When combined, the structured metadata and unstructured narratives provide complementary perspectives: metadata facilitates quantitative aggregation and trend analysis, while narratives reveal the underlying stories and latent factors that contribute to accidents [15][3]. The process of analyzing accident narratives begins with data pre-processing, which involves cleaning the text, removing irrelevant information, and standardizing terminology. Advanced NLP models, such as pre-trained SBERT (Sentence-BERT), are then employed to generate dense vector representations, sentence embeddings, that capture the semantic meaning of each narrative [2][11]. These embeddings are high-dimensional, often exceeding a thousand dimensions, reflecting the complexity and variability of the language used in accident reports [2]. The high dimensionality poses challenges for direct analysis and visualization, necessitating the use of dimensionality reduction techniques such as UMAP (Uniform Manifold Approximation and Projection). UMAP effectively projects the embeddings into a lower-dimensional space, preserving the local and global structure of the data, which is crucial for subsequent clustering and pattern discovery [6][2]. Clustering algorithms, particularly k-means, are then applied to the reduced embeddings to group similar accident narratives together. This approach uncovers latent patterns and thematic structures that may not be apparent from metadata alone [15][2]. For instance, clusters may correspond to specific types of incidents, recurring causal factors, or shared environmental conditions. The integration of metadata with these clusters allows for the validation and interpretation of the discovered patterns, as metadata variables can be analyzed within and across clusters to identify distinguishing features. Rose et al. [15] indicate that even within well-defined clusters, there can be meaningful subdivisions, each characterized by distinct metadata values, underscoring the importance of considering both narrative and structured data in accident analysis. The combination of narrative embeddings and metadata not only enhances the granularity of accident analysis but also supports the development of predictive models and targeted interventions. For example, content analysis of accident scenarios, when supported by software tools and models, has been shown to reveal previously unnoticed trends and inform prevention strategies in mining safety research [1]. The use of advanced NLP and machine learning techniques further enables the extraction of hidden correlations and thematic structures from large-scale datasets, as demonstrated in studies of both mining and aviation safety [23][1]. In summary, the nature of accident narratives and metadata is fundamentally complementary. Narratives provide depth and context, capturing the complexity of real-world incidents, while metadata offers structure and comparability. The integration of these data types, facilitated by modern NLP and clustering methodologies, represents a significant advancement in the analysis of mining accident data, enabling deeper insights and more effective safety interventions [3][2].

4.2 Enrichment Strategies for Textual Data

4.2.1 Merging Metadata with Narratives

Merging metadata with accident narratives represents a crucial enrichment strategy for mining accident analysis, as it enables the integration of structured and unstructured data to enhance the semantic representation of incidents. Accident narratives, typically unstructured textual descriptions, capture nuanced contextual information about events, while metadata variables, such as accident type, location, time, and body part affected, provide structured categorical or numerical descriptors that are essential for comprehensive analysis [2][19]. By combining these two data modalities, the resulting dataset encapsulates both the semantic depth of narrative text and the categorical precision of metadata, thereby supporting more robust downstream analytical tasks. The process begins with the generation of sentence embeddings from accident narratives using pre-trained models such as SBERT. These embeddings encode the semantic content of the narratives into high-dimensional vector representations, capturing relationships and contextual information that are not readily accessible through traditional keyword-based approaches [2][3]. However, narrative text alone may omit critical structured details, such as the specific nature of the accident or the part of the body involved, which are systematically recorded in metadata fields. Therefore, concatenating or otherwise integrating metadata variables with the narrative embeddings enriches the feature space, allowing the model to leverage both the latent semantic information and explicit categorical attributes. This enrichment is particularly valuable in the context of mining accident analysis, where the complexity of incidents often arises from the interplay of multiple factors. For example, the inclusion of metadata such as accident classification codes, event titles, and body parts affected, coded according to standardized taxonomies like the Occupational Injury and Illness Classification Manual, enables the model to discern patterns that may be obscured in narrative text alone [2]. The authors of [19] indicate that preprocessing steps, such as the addition of prompt words or tokens to encode metadata, can facilitate the model's understanding of discrete accident elements, further enhancing the integration of structured and unstructured data. The combined feature vectors, comprising both narrative embeddings and metadata variables, are then subjected to dimensionality reduction techniques such as UMAP. This step is essential for visualizing and clustering high-dimensional data, as it preserves the local and global structure of the original feature space while reducing computational complexity. UMAP, in particular, has demonstrated superior performance in maintaining cluster integrity compared to linear methods like PCA, especially when dealing with complex, high-dimensional representations derived from enriched data. The reduced representations are subsequently analyzed using clustering algorithms such as k-means, which can identify latent patterns and groupings within the enriched dataset, revealing trends and risk factors that may not be apparent through traditional analysis [6]. Integrating metadata with narrative embeddings also supports advanced natural language processing (NLP) and machine learning workflows. For instance, the use of special tokens to demarcate metadata fields, as described in, allows transformer-based models to capture the semantic interplay between structured and unstructured components. The cosine similarity between the resulting embeddings can then be used to measure the relatedness of incidents, facilitating tasks such as similarity search, classification, and anomaly detection [2][3]. This approach aligns with recent advances in accident analysis, where the fusion of textual and structured data has been shown to improve the identification of causal factors and the prediction of future incidents [1][23]. Furthermore, the integration of metadata and narratives addresses challenges associated with synonymy and lexical variation in accident reports. As outlined in [8], the development of reliable lexicons and the inclusion of metadata can help disambiguate terms and ensure consistent interpretation of incident descriptions, particularly in multilingual or domain-specific contexts. This is especially relevant in mining, where technical terminology and reporting practices may vary across sites and jurisdictions. By enriching narrative embeddings with metadata, researchers can construct a more holistic and semantically rich representation of mining accidents. This integrated approach not only enhances the interpretability and accuracy of clustering and classification models but also provides deeper insights into the underlying causes and dynamics of accidents, supporting more effective prevention and intervention strategies [6][2][19].

4.2.2 Semantic Enhancement and Contextualization

Semantic enhancement and contextualization are essential for extracting meaningful information from unstructured accident narratives, especially when these narratives are integrated with structured metadata. The use of pre-trained SBERT models to generate sentence embeddings represents a significant advancement in capturing the nuanced semantics of accident reports. SBERT, when fine-tuned on natural language inference data, produces sentence representations that outperform other state-of-the-art embedding methods in terms of semantic textual similarity, enabling more accurate and context-aware analysis of textual data [17]. This semantic richness is crucial for mining accident narratives, where subtle differences in language can indicate distinct causal factors or outcomes. The integration of these embeddings with metadata variables from structured sources, such as the MSHA accident database, further contextualizes the textual information. By combining semantic vectors with structured attributes, the analysis benefits from both the depth of narrative context and the precision of categorical or numerical data. This dual enrichment strategy allows for a more comprehensive understanding of accident causation and trends, as it leverages the strengths of both data types [3][21]. For instance, metadata such as accident type, location, or time can be used to anchor the semantic content of narratives, facilitating more targeted clustering and pattern recognition. Dimensionality reduction techniques, such as UMAP, play a critical role in this process by transforming high-dimensional semantic embeddings into lower-dimensional spaces that preserve the essential structure of the data. This step is vital for effective clustering, as it enables the identification of latent patterns and relationships that may not be apparent in the original high-dimensional space. The resulting clusters can reveal common themes, recurring hazards, or emerging trends in accident data, providing actionable insights for risk mitigation and safety management. The process of semantic enhancement is further strengthened by the use of domain-specific lexicons and the refinement of language models tailored to the context of mining accidents. Constructing such lexicons and adapting language models to the specific terminology and phraseology of the mining industry can improve the reliability of extracted themes and factors. However, even in the absence of fully automated domain adaptation, manual approaches that combine keyword analysis with semantic analysis have proven effective in identifying relevant clusters and themes within accident narratives [3]. Automated semantic enrichment also addresses the limitations of traditional, expert-driven analyses, which are often subjective and inconsistent. By leveraging NLP-based techniques, the extraction and contextualization of hazards and precursors become more objective and reproducible, reducing the reliance on individual judgment and enabling scalable analysis across large datasets [21][13]. This objectivity is particularly valuable in high-stakes domains such as mining, where timely and accurate identification of risks can have significant safety implications. The use of pre-trained word embeddings, such as those provided by SBERT, is integral to modern NLP systems and offers substantial improvements over embeddings learned from scratch [5]. These embeddings capture both syntactic and semantic relationships, allowing for a more nuanced representation of accident narratives. When combined with structured metadata, they enable a holistic approach to accident analysis that transcends the limitations of purely rule-based or manual methods [21]. Furthermore, the clustering of semantically enriched and contextually grounded embeddings facilitates the discovery of patterns that may inform prescriptive analyses and rule extraction. Such analyses can guide the development of targeted interventions and safety protocols, moving beyond descriptive statistics to actionable recommendations [27]. The integration of semantic enhancement, contextualization, and clustering thus represents a comprehensive framework for advancing accident analysis in the mining sector. The application of these techniques is not limited to mining; similar approaches have demonstrated value in other safety-critical domains, such as aviation, where topic modeling and sentiment analysis have been used to systematically uncover latent themes in accident reports [13][15]. The scalability and objectivity of these methods make them particularly suited for large-scale safety data analysis, where manual review is impractical. In summary, semantic enhancement and contextualization, achieved through the integration of SBERT-based embeddings, metadata enrichment, and dimensionality reduction, provide a robust foundation for advanced clustering and pattern discovery in accident analysis. This approach enables a deeper, more context-aware understanding of accident causation and supports the development of effective risk mitigation strategies [17][3][21][13].

4.3 Benefits and Challenges of Combined Data Analysis

The integration of structured metadata and unstructured accident narratives in mining safety analysis offers substantial benefits, yet it also introduces several challenges that must be addressed to maximize analytical value. By leveraging pre-trained SBERT models to generate sentence embeddings from narrative text, the approach captures nuanced semantic information that traditional structured data alone cannot provide. This semantic enrichment enables the identification of subtle patterns and relationships within accident reports, which are often lost when relying solely on categorical or numerical metadata [28]. The combination of these embeddings with structured variables, such as those mandated by MSHA reporting requirements, creates a comprehensive dataset that reflects both the context and the specifics of each incident [25][29]. One of the primary benefits of this combined analysis is the ability to uncover complex, multifactorial trends in accident causation and outcomes. Structured data, such as injury type, location, and occupation, provides a standardized framework for comparison across incidents [25]. However, unstructured narratives often contain critical details about the sequence of events, contributing factors, and environmental conditions that are not captured in fixed fields [8][30]. By integrating these two data types, researchers can perform more holistic analyses, leading to richer insights into the underlying causes of accidents and more targeted recommendations for safety interventions [29]. The use of advanced text mining and natural language processing techniques, such as tokenization, stopword removal, and contextual embedding generation, further enhances the quality and interpretability of the unstructured data [23][28]. Dimensionality reduction techniques, such as UMAP, play a crucial role in managing the high dimensionality of combined datasets. These methods facilitate the visualization and clustering of complex data, making it possible to identify distinct patterns and groupings that may correspond to specific risk factors or accident typologies. The application of k-means clustering to the reduced embeddings enables the discovery of latent structures within the data, supporting the development of data-driven safety programs and interventions. The elbow method, for instance, assists in determining the optimal number of clusters, ensuring that the analysis remains both interpretable and actionable [6]. Despite these advantages, several challenges arise when integrating structured and unstructured data. One significant issue is the inherent imbalance and sparsity present in real-world accident datasets. Minority classes, such as rare accident types or infrequent causes, are often underrepresented, which can hinder the performance of machine learning algorithms and bias the results toward more common events. This imbalance necessitates careful preprocessing and, in some cases, the use of specialized techniques to ensure that rare but critical patterns are not overlooked [9][27]. Another challenge lies in the preprocessing and harmonization of heterogeneous data sources. Unstructured narratives require extensive cleaning, including tokenization, lowercasing, punctuation and stopword removal, and the elimination of irrelevant elements such as URLs [23]. The development of reliable lexicons and the identification of domain-specific keywords are essential for accurate text mining and entity recognition [8][30]. Furthermore, the training and fine-tuning of word and sentence embedding models, such as those based on word2vec or SBERT, demand substantial computational resources and domain expertise to ensure that the resulting representations are both meaningful and generalizable. Interpretability is also a concern when employing advanced machine learning and deep learning models for combined data analysis. While these models can capture complex relationships, their decision-making processes are often opaque, making it difficult to translate findings into actionable safety recommendations [22]. The integration of attention mechanisms, as seen in self-attention models, can partially address this issue by highlighting the most influential words or phrases in accident narratives, thereby improving transparency and trust in the analytical outcomes [19]. Visualization and communication of results present additional challenges. High-dimensional embeddings and cluster assignments must be translated into intuitive plots and summaries that can be understood by safety managers and other stakeholders. Tools such as matplotlib facilitate the creation of informative visualizations, but careful design is required to ensure that key findings are accessible and actionable. The process of mapping extracted hazards and accident features to visual representations supports the identification of priority areas for intervention and training [21][25]. The integration of structured and unstructured data in mining accident analysis thus offers a powerful framework for advancing safety research. It enables the extraction of deeper insights from complex datasets, supports the identification of emerging trends, and informs the development of targeted safety programs. However, realizing these benefits requires addressing challenges related to data imbalance, preprocessing, model interpretability, and effective communication of results [8][25][23][9].

5 Sentence Embeddings for Accident Narratives

5.1 Overview of Sentence-BERT (SBERT)

Sentence-BERT (SBERT) represents a significant advancement in the field of sentence embeddings, specifically designed to address the limitations of traditional BERT models in generating semantically meaningful vector representations for sentences. Unlike the original BERT architecture, which is primarily optimized for token-level tasks and requires computationally expensive pairwise comparisons for sentence similarity, SBERT introduces a modification that enables efficient and effective computation of sentence-level embeddings suitable for a wide range of downstream tasks. The core innovation of SBERT lies in its use of a siamese or triplet network structure, where two or more sentences are passed independently through the same BERT-based encoder to produce fixed-size embeddings. These embeddings can then be directly compared using cosine similarity or other distance metrics, making SBERT particularly well-suited for tasks such as semantic textual similarity, clustering, and information retrieval. The authors of indicate that SBERT achieves substantial improvements in sentiment analysis and semantic similarity tasks, outperforming previous models like InferSent and Universal Sentence Encoder on several benchmarks. This is attributed to SBERT's ability to capture nuanced semantic information at the sentence level, which is essential for applications involving complex narrative data such as accident reports. SBERT's architecture leverages the strengths of BERT's deep contextualized representations while introducing pooling strategies, such as mean or max pooling, over the token embeddings to generate a single vector for each sentence. While previous research found that max pooling could be beneficial for certain models like InferSent, SBERT demonstrates robust performance with mean pooling, balancing computational efficiency and representational quality. This design choice is particularly advantageous when processing large-scale datasets, as it allows for rapid computation of embeddings for millions of sentences, a necessity in domains like accident analysis where narrative data is abundant. The evaluation of SBERT embeddings is often conducted using frameworks such as SentEval, which assess the quality of sentence representations across various classification and similarity tasks. Results from show that SBERT consistently delivers high accuracy and improved performance on sentiment and semantic similarity datasets, with the exception of specific cases like the TREC dataset, where Universal Sentence Encoder may have a slight edge. Nevertheless, the overall gains in both accuracy and computational speed make SBERT a compelling choice for embedding accident narratives. In the context of mining accident analysis, the ability of SBERT to generate dense, semantically rich embeddings from narrative text enables the integration of unstructured textual information with structured metadata. This fusion supports advanced analytical workflows, such as dimensionality reduction with UMAP and subsequent clustering with k-means, to uncover latent patterns and trends in accident data. The computational efficiency of SBERT, as highlighted in [17], ensures that even large-scale narrative datasets can be processed feasibly, facilitating comprehensive and scalable analyses. Furthermore, the flexibility of SBERT embeddings extends to their use in various downstream applications, including classification, clustering, and retrieval, without the need for extensive architectural modifications. This adaptability is crucial for research scenarios where the nature of the analysis may evolve or where multiple analytical perspectives are required [18]. The integration of SBERT with other machine learning techniques, such as gradient boosting models or neural network-based cluster explanations, further enhances the interpretability and actionable insights derived from accident narrative data [7][15]. By leveraging SBERT, researchers can move beyond surface-level keyword extraction or basic thematic modeling, which often suffer from issues like overfitting or inadequate representation of information diversity [3]. Instead, SBERT provides a robust foundation for capturing the full semantic content of accident narratives, supporting more nuanced and data-driven safety analyses. This capability is particularly valuable when combined with metadata variables, as it allows for a holistic understanding of the factors contributing to mining accidents and supports the development of targeted prevention strategies.

5.2 Comparative Analysis with Other Embedding Techniques

A comparative analysis of sentence embeddings generated by pre-trained SBERT models and other embedding techniques reveals significant differences in semantic representation, computational efficiency, and suitability for mining accident narrative analysis. SBERT, or Sentence-BERT, is specifically designed to produce semantically meaningful sentence-level embeddings, mapping entire sentences into fixed-length vectors that capture contextual and relational information between words within a sentence. This contrasts with traditional word embedding methods such as Word2Vec and GloVe, which focus on word-level representations and require additional pooling strategies to aggregate word vectors into sentence or document representations [3][17]. The granularity of representation is a fundamental distinction. Word embeddings, as outlined by Zheng Maa et al., map individual words into a vector space, resulting in a loss of inter-word dependencies when aggregated for sentence-level tasks. In contrast, sentence embeddings like those produced by SBERT directly encode the semantics of the entire sentence, preserving contextual nuances that are critical for understanding complex accident narratives [3]. This property is particularly advantageous when analyzing unstructured text data, such as accident reports, where the meaning often depends on the interplay of multiple words and phrases. Computational efficiency is another important consideration. Reimers and Gurevych demonstrate that SBERT achieves high computation speed when generating sentence embeddings, outperforming average GloVe embeddings and other models such as InferSent and Universal Sentence Encoder in large-scale scenarios. This efficiency is crucial when processing millions of accident narratives, as required in mining safety research. The ability to rapidly compute embeddings without sacrificing semantic richness enables scalable analysis and supports downstream tasks such as clustering and trend identification [17]. Traditional word embedding techniques, such as Word2Vec, require careful parameter tuning to balance computational cost and embedding quality. For instance, the choice of vector dimensionality, window size, and negative sampling parameters can significantly impact both the training time and the expressiveness of the resulting embeddings [22]. Fan Zhang [11] notes that lower word vector dimensionality reduces computation time, while higher-order n-gram settings increase the time cost but may capture more complex relationships. However, even with optimal settings, word embeddings lack the sentence-level semantic integration that SBERT provides. Document embeddings, which aggregate information at the document level, offer another alternative. However, for tasks focused on sentence-level analysis, such as extracting patterns from individual accident narratives, document embeddings may be too coarse, potentially obscuring important details present at the sentence level [3]. Sentence embeddings thus strike a balance between granularity and semantic depth, making them particularly suitable for the analysis of narrative accident data. The effectiveness of sentence embeddings is further highlighted when considering their application to downstream tasks such as clustering. The high-dimensional semantic vectors produced by SBERT can be effectively reduced using techniques like UMAP, facilitating the identification of patterns and trends through clustering algorithms such as k-means. This approach enables the discovery of latent structures in accident data that may not be apparent through traditional statistical or keyword-based analyses [17][3]. In contrast, keyword-based methods, while useful for hazard identification, often rely on simple token matching and lack the ability to capture deeper semantic relationships within the text [21]. Datasets such as SICK, SNLI, and MultiNLI have been instrumental in benchmarking the performance of various embedding techniques, demonstrating that models leveraging sentence-level embeddings consistently outperform those based solely on word-level representations in tasks involving semantic relatedness and entailment. Karlo Babić et al. [18] emphasize that while word-level input is standard, subword- and sentence-level approaches can yield superior results, particularly in complex classification tasks. The integration of SBERT-based sentence embeddings with metadata variables from structured accident data, as implemented in this research, represents a methodological advancement over traditional approaches that rely primarily on statistical analysis or manual coding of accident causes [31][25]. Xu et al. [31] highlight the limitations of manual statistical analysis, which often considers only a single primary cause per accident and is labor-intensive. By contrast, the proposed framework leverages the semantic richness of sentence embeddings to capture multifaceted accident scenarios, enabling more nuanced and comprehensive analysis. In summary, the comparative analysis underscores the advantages of SBERT-based sentence embeddings over traditional word and document embedding techniques in the context of mining accident narrative analysis. SBERT offers superior semantic representation at the sentence level, computational efficiency, and enhanced capability for pattern discovery when combined with dimensionality reduction and clustering methods [17][3][18][31]. This integrated approach provides deeper insights into accident causation and trends, supporting more effective safety interventions and risk management strategies.

5.3 Semantic Richness in Accident Reporting

Semantic richness in accident reporting is fundamentally enhanced by the use of advanced sentence embedding techniques, which capture nuanced contextual and semantic information from narrative data. Traditional accident reports, while valuable, often present information in a structured or categorical format, limiting the depth of insight that can be extracted from the underlying narratives [8][32]. By leveraging pre-trained models such as SBERT, it becomes possible to encode entire accident narratives into dense vector representations that encapsulate not only the explicit content but also the subtle relationships and latent meanings present in the text [18][5]. These embeddings are inherently language-agnostic and task-independent, allowing for a more generalizable and robust analysis across diverse datasets and reporting styles [18]. The semantic representations generated by SBERT and similar models are particularly adept at capturing sentence-level meaning, which is crucial for understanding the complex interplay of factors described in accident narratives [3][5]. Unlike word-level embeddings that focus on isolated terms, sentence embeddings aggregate contextual cues, syntactic structure, and discourse-level information, resulting in a holistic representation of each report [3]. This enables the identification of patterns and trends that may not be apparent through manual review or simpler text analysis methods [13][10]. Integrating these semantic-rich embeddings with metadata variables from structured accident databases, such as those provided by the Mine Safety and Health Administration (MSHA), further enriches the analytical framework. Metadata, such as accident type, location, time, and worker demographics, provides essential context that, when combined with narrative embeddings, allows for a multidimensional exploration of accident causality and risk factors [25][8]. The fusion of narrative and structured data supports a more comprehensive understanding of the circumstances surrounding each incident, facilitating the discovery of latent clusters and emergent themes within the data [10][13]. Dimensionality reduction techniques, such as UMAP, play a critical role in preserving the semantic relationships encoded in high-dimensional embeddings while making the data amenable to visualization and clustering. The reduction process maintains the integrity of the latent semantic space, ensuring that similar narratives remain proximate in the reduced space, which is essential for effective pattern recognition and cluster formation [10]. This approach enables the application of clustering algorithms like k-means to group accident reports based on shared semantic characteristics, revealing underlying structures and trends that inform targeted safety interventions [19][9]. The use of deep learning architectures, including BiLSTM and CNN-based models, has demonstrated superior performance in extracting and classifying semantic features from accident narratives compared to traditional methods [22][20]. These models, particularly when augmented with attention mechanisms, are capable of discerning critical information and highlighting salient aspects of the text that contribute to accident causation and severity [22]. The ability to automatically learn and prioritize relevant features from unstructured data represents a significant advancement in the field of accident analysis [20]. Furthermore, the interpretability of semantic-rich embeddings is enhanced by their compatibility with downstream tasks such as topic modeling, risk factor identification, and predictive analytics [13][9]. For instance, topic modeling techniques applied to embedded narratives can extract dominant themes and recurring issues, providing actionable insights for safety management and policy development [13]. The integration of semantic embeddings with machine learning models, such as random forests or gradient boosting, enables the prediction of key safety outcomes, including injury type and severity, based on the narrative content and associated metadata [9][25]. The transformation of accident narratives into semantically rich, low-dimensional representations thus marks a paradigm shift in accident reporting and analysis. It moves the field beyond the limitations of manual coding and categorical analysis, offering a scalable and data-driven approach to uncovering the complex, multifactorial nature of workplace accidents [32][8]. This methodological innovation not only enhances the granularity and depth of accident investigations but also supports the development of more effective, evidence-based safety interventions [32][25].

6 Dimensionality Reduction for High-Dimensional Text Embeddings

6.1 Need for Dimensionality Reduction in Text Mining

The necessity for dimensionality reduction in text mining arises from the inherent complexity and high dimensionality of textual data representations. When employing advanced language models such as SBERT to generate sentence embeddings, each narrative is mapped to a high-dimensional vector space, often with hundreds or thousands of dimensions. This high dimensionality is a direct consequence of capturing nuanced semantic relationships between words and phrases, as similar concepts, such as ‘scaffold’ and ‘ladder’ or ‘arc flash’ and ‘electrical fault’, are positioned closely in this space despite their lexical differences [11]. However, the resulting vector space is not only computationally intensive to process but also challenging to visualize and interpret, making downstream analysis such as clustering or pattern recognition less tractable. The curse of dimensionality is a well-documented phenomenon in high-dimensional data analysis, where the volume of the space increases so rapidly that the available data become sparse, and traditional distance metrics lose their discriminative power. This sparsity can hinder the performance of machine learning algorithms, as models may struggle to generalize or identify meaningful patterns. Furthermore, high-dimensional spaces often contain redundant or irrelevant features, which can introduce noise and obscure the underlying structure of the data. As a result, reducing the dimensionality of text embeddings is essential to mitigate these issues and enhance the efficiency and effectiveness of subsequent analytical tasks. Dimensionality reduction techniques serve to project high-dimensional data into a lower-dimensional space while preserving as much of the intrinsic structure and information as possible. The concept of intrinsic dimension refers to the minimum number of variables required to represent the data without significant loss of information [33]. By identifying and retaining only the most informative features, dimensionality reduction not only alleviates computational burdens but also facilitates the discovery of latent patterns and relationships within the data. For example, methods such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) have been widely adopted for this purpose, with PCA offering a linear approach that retains variance and t-SNE providing a non-linear mapping suitable for visualization [15]. UMAP, as utilized in this research, extends these capabilities by efficiently capturing both local and global data structures, making it particularly well-suited for high-dimensional text embeddings. The selection of an appropriate dimensionality reduction method is itself a non-trivial challenge, as different techniques may yield varying results depending on the nature of the data and the intended application [33]. It is crucial to balance the trade-off between reducing dimensionality and preserving the semantic richness of the original embeddings. In text mining, where semantic relationships are paramount, methods that maintain neighborhood integrity and global structure are preferred. The authors of [15] indicate that combining dimensionality reduction with clustering algorithms can minimize information loss and enhance the interpretability of complex datasets. Moreover, the integration of metadata variables with text embeddings further increases the dimensionality and heterogeneity of the data, amplifying the need for effective reduction strategies [7]. Without dimensionality reduction, clustering algorithms such as k-means may perform suboptimally due to the high-dimensional noise and the difficulty in distinguishing meaningful groupings. By projecting the data into a lower-dimensional space, these algorithms can more readily identify coherent clusters that correspond to distinct patterns or trends in accident narratives [6][15]. Latent Semantic Analysis (LSA) exemplifies an early approach to dimensionality reduction in text mining, where a term-document matrix is decomposed to uncover latent semantic structures and associations among words [33]. This foundational idea underpins more recent advances in embedding-based representations, highlighting the enduring importance of dimensionality reduction for extracting actionable insights from textual data. In summary, dimensionality reduction is indispensable in text mining applications that leverage high-dimensional embeddings. It addresses computational challenges, mitigates the curse of dimensionality, and enhances the interpretability and effectiveness of clustering and other analytical techniques. The careful application of reduction methods such as UMAP, in conjunction with clustering, enables researchers to uncover deeper insights from complex, semantically rich datasets that would otherwise remain obscured in high-dimensional spaces [11][15][33].

6.2 UMAP: Principles and Applications

6.2.1 Comparison with Other Techniques (e.g., t-SNE, PCA)

A critical aspect of dimensionality reduction for high-dimensional text embeddings is the selection of an appropriate technique that preserves the semantic structure of the data while enabling effective downstream analysis such as clustering. UMAP, t-SNE, and PCA are among the most widely adopted methods, each with distinct principles and performance characteristics. UMAP (Uniform Manifold Approximation and Projection) is recognized for its ability to maintain both local and global data structures during dimensionality reduction, which is particularly advantageous when working with complex, high-dimensional embeddings such as those generated by SBERT. In comparative studies, UMAP has demonstrated competitive performance, especially when the dimensionality reduction ratio is substantial, such as reducing to 1/100 of the original dimensions. However, it is important to note that UMAP does not always yield the best clustering outcomes for every dataset. For instance, when applied to the Coil 20 dataset, the k-NN accuracy without any dimensionality reduction was 0.956, indicating that UMAP did not surpass the original high-dimensional representation in terms of clustering performance. This suggests that while UMAP is effective in many scenarios, its performance can be dataset-dependent and may not universally outperform other methods. In contrast, t-SNE (t-Distributed Stochastic Neighbor Embedding) is particularly effective at preserving local neighborhood structures, making it a popular choice for visualizing high-dimensional data in two or three dimensions. When the dimensionality is reduced to three, t-SNE has been observed to outperform UMAP and PCA in terms of visual separation and clustering quality. The non-linear nature of t-SNE allows it to capture complex relationships within the data, which is beneficial for exploratory analysis and pattern discovery in accident narratives. However, t-SNE is computationally intensive and less scalable to very large datasets, which can limit its applicability in large-scale mining accident analyses [6][15]. PCA (Principal Component Analysis), on the other hand, is a linear technique that projects data onto orthogonal axes of maximum variance. It is widely used for its simplicity and computational efficiency, especially in exploratory analyses of high-dimensional data. PCA is effective when the underlying data structure is approximately linear and when the goal is to retain as much variance as possible in the reduced dimensions. However, PCA may not capture non-linear relationships inherent in semantic-rich embeddings, potentially leading to suboptimal clustering or loss of important information. Despite these limitations, PCA remains a valuable baseline and is often used in conjunction with other methods to assess the robustness of dimensionality reduction outcomes. The elbow method is commonly employed to determine the optimal number of clusters after dimensionality reduction, regardless of the technique used. By plotting the within-cluster sum of squares (WCSS) against the number of clusters, the inflection point can be identified as the optimal cluster count, providing a systematic approach to clustering analysis. This method is applicable across UMAP, t-SNE, and PCA, ensuring consistency in evaluating clustering performance. In practical applications, the choice between UMAP, t-SNE, and PCA should be guided by the specific characteristics of the dataset, the computational resources available, and the analytical objectives. For example, UMAP and t-SNE are preferable when non-linear relationships and local structure preservation are critical, while PCA is suitable for rapid, linear dimensionality reduction. The integration of these techniques with advanced clustering algorithms, such as k-means, enables the identification of meaningful patterns and trends in mining accident data, offering insights that extend beyond traditional analysis methods [6][15]. Furthermore, the interpretability of the reduced dimensions and the resulting clusters is enhanced when these techniques are combined with metadata variables and semantic-rich embeddings. This integrated framework supports a more nuanced understanding of accident narratives, facilitating the discovery of latent patterns that may inform safety interventions and policy development [6][15]. The comparative analysis of UMAP, t-SNE, and PCA underscores the importance of methodological flexibility and empirical validation in high-dimensional text analytics. By leveraging the strengths of each technique and critically assessing their limitations, researchers can optimize the extraction of actionable knowledge from complex accident datasets.

6.2.2 Suitability for Accident Analysis Data

The suitability of UMAP for accident analysis data, particularly when working with high-dimensional sentence embeddings derived from accident narratives, is grounded in its ability to effectively capture both local and global data structures in a reduced-dimensional space. Accident narratives, when encoded using pre-trained SBERT models, yield embeddings that are semantically rich but inherently high-dimensional, often exceeding several hundred dimensions. This high dimensionality, while beneficial for capturing nuanced semantic relationships, poses significant challenges for downstream analysis such as clustering, due to the curse of dimensionality and computational inefficiency [33]. UMAP addresses these challenges by providing a nonlinear dimensionality reduction technique that is well-suited for complex, structured data such as text embeddings. Unlike linear methods such as PCA, which may fail to preserve nonlinear relationships present in the data, UMAP is designed to maintain both the local neighborhood structure and the broader global arrangement of data points. This is particularly advantageous for accident analysis, where subtle semantic differences between narratives can be critical for distinguishing between different types of incidents or underlying risk factors. Hozumi et al. [6] state that UMAP, along with t-SNE, is capable of projecting high-dimensional data onto a low-dimensional manifold while preserving essential data structures, which is crucial for meaningful clustering. The integration of metadata variables from MSHA accident data with SBERT embeddings further increases the dimensionality and heterogeneity of the dataset. UMAP's flexibility in handling mixed-type data and its scalability to large datasets make it a practical choice for this context [33]. The method's ability to reduce dimensionality without significant loss of information enables the application of clustering algorithms such as k-means, which are sensitive to the geometry of the input space. Empirical results demonstrate that UMAP can enhance the performance of clustering methods, as evidenced by improved accuracy in k-means clustering tasks on benchmark datasets. However, it is important to note that the effectiveness of UMAP may vary depending on the specific characteristics of the dataset, such as the intrinsic dimensionality and the distribution of classes. For instance, while UMAP outperformed PCA in clustering tasks on certain datasets, its performance was not universally superior across all evaluation metrics or datasets. Another consideration is the interpretability of the reduced dimensions. UMAP's nonlinear transformations can sometimes make it challenging to directly interpret the resulting axes in terms of original features. Nevertheless, for exploratory analysis and pattern discovery in accident data, the primary objective is often to reveal clusters or trends that are not apparent in the original high-dimensional space, rather than to interpret individual dimensions [6]. The ability of UMAP to reveal such latent structures aligns well with the goals of accident analysis, where uncovering hidden patterns can inform targeted interventions and safety improvements. The selection of dimensionality reduction technique is a critical step in the analysis pipeline. As outlined in [33], the choice must be tailored to the nature of the data and the intended analytical objectives. UMAP's suitability for accident analysis data is further supported by its robustness to noise and outliers, which are common in real-world narrative datasets. The method's parameterization, such as the number of neighbors and minimum distance, allows for fine-tuning to balance the preservation of local versus global structures, offering flexibility to adapt to the specific requirements of mining accident data. In summary, UMAP provides a powerful and adaptable framework for reducing the dimensionality of high-dimensional, semantically rich accident narrative embeddings, especially when combined with metadata. Its ability to preserve meaningful data structures facilitates effective clustering and pattern recognition, which are essential for extracting actionable insights from complex accident datasets [6][33].

6.3 Preserving Semantic Structure in Low-Dimensional Spaces

Preserving the semantic structure of high-dimensional text embeddings during dimensionality reduction is essential for meaningful downstream analysis, such as clustering and visualization. When sentence embeddings, such as those generated by pre-trained SBERT models, are combined with metadata and projected into a lower-dimensional space, the challenge lies in maintaining both local and global relationships that encode semantic similarity and contextual information [6]. Dimensionality reduction techniques, particularly those designed for nonlinear manifolds, are specifically developed to address this challenge by mapping complex, high-dimensional data into a space where intrinsic semantic relationships are retained as much as possible [33]. Linear methods like Principal Component Analysis (PCA) often fail to capture nonlinear dependencies inherent in semantic-rich embeddings, leading to a loss of meaningful structure. In contrast, nonlinear algorithms such as t-SNE and UMAP are capable of preserving both local neighborhoods and broader global patterns, which is crucial for text data where subtle semantic differences can be significant [6][15]. UMAP, in particular, constructs a topological representation of the data, aiming to maintain the underlying manifold structure during the reduction process. This approach ensures that sentences with similar meanings remain close together in the low-dimensional space, while dissimilar sentences are mapped further apart, thus preserving the semantic integrity of the original embeddings [33]. The process of dimensionality reduction can be mathematically described as finding a mapping $f : \mathbb{R}^M \to \mathbb{R}^k$ with $k \ll M$, such that the pairwise relationships between data points, often modeled as probability distributions or distances, are preserved as faithfully as possible. For example, t-SNE computes pairwise similarities in the high-dimensional space and seeks a low-dimensional embedding where these similarities are reflected in the spatial arrangement of points [6]. UMAP extends this idea by leveraging both local and global data structure, using techniques from algebraic topology to construct a fuzzy simplicial set that encodes the data's manifold, and then optimizing a low-dimensional representation that best preserves this structure [33]. The preservation of semantic structure is not only a theoretical concern but has practical implications for clustering and pattern discovery. When the low-dimensional representation accurately reflects the semantic relationships present in the original embeddings, clustering algorithms such as k-means can more effectively identify meaningful groups and trends within the data [8][15]. This is particularly important in accident analysis, where subtle differences in narrative descriptions and associated metadata can correspond to distinct types of incidents or risk factors. By maintaining the semantic structure, the reduced space enables the discovery of patterns that would otherwise be obscured by the curse of dimensionality [33]. Furthermore, the integration of metadata variables with sentence embeddings enriches the feature space, providing additional context that can enhance the preservation of semantic relationships during reduction [11]. However, this also increases the dimensionality, making the choice of an effective reduction technique even more critical. Nonlinear methods like UMAP are well-suited to handle such complex, heterogeneous data, as they are designed to capture both the semantic and contextual nuances embedded in the high-dimensional representations [6]. The authors of [33] indicate that dimensionality reduction transforms high-dimensional data into a lower-dimensional form while striving to retain the original meaning and relationships. This transformation mitigates the curse of dimensionality, making the data more amenable to analysis and visualization. The ability to visualize clusters and trends in two or three dimensions, while still reflecting the underlying semantic structure, is a significant advantage for interpretability and actionable insights [6][33]. In summary, the preservation of semantic structure in low-dimensional spaces is achieved through the careful selection and application of nonlinear dimensionality reduction techniques, which are capable of maintaining both local and global relationships present in semantic-rich embeddings. This enables more effective clustering, visualization, and interpretation of complex text data, ultimately supporting deeper insights into patterns and trends within accident narratives and their associated metadata [6][33][15].

7 Clustering Techniques for Accident Pattern Discovery

7.1 Clustering Approaches in Text Mining

Clustering approaches in text mining have evolved significantly with the advent of advanced natural language processing (NLP) models and dimensionality reduction techniques. Traditional clustering methods, such as k-means, have long been employed to group similar data points based on feature similarity, but their effectiveness is often limited when applied directly to high-dimensional and semantically complex text data. The integration of pre-trained language models, such as BERT and its variants, has transformed the representation of textual information, enabling the extraction of dense, context-aware embeddings that capture nuanced semantic relationships between sentences and documents. These embeddings, when combined with relevant metadata, provide a rich feature space for subsequent clustering analysis. The use of sentence embeddings generated by models like SBERT allows for the encoding of accident narratives into high-dimensional vectors that encapsulate both syntactic and semantic information [17]. This approach addresses the limitations of traditional bag-of-words or TF-IDF representations, which often fail to capture the contextual meaning inherent in narrative data. By leveraging these embeddings, clustering algorithms can more effectively discern subtle patterns and groupings within the data, which is particularly valuable in domains such as accident analysis where the underlying causes and circumstances are often complex and multifaceted [16][12]. However, the high dimensionality of sentence embeddings poses challenges for clustering algorithms, especially those like k-means, which are sensitive to the curse of dimensionality and may struggle to identify meaningful clusters in such spaces. To mitigate these issues, dimensionality reduction techniques such as UMAP (Uniform Manifold Approximation and Projection) are employed to project the high-dimensional embeddings into a lower-dimensional space while preserving the local and global structure of the data. UMAP has demonstrated stable performance across various reduction ratios and is particularly advantageous for preparing data for clustering, as it enhances the separability of clusters and improves the reliability of subsequent analyses. The authors of indicate that while t-SNE may outperform UMAP in certain scenarios, UMAP offers greater stability and reliability, making it a preferred choice for clustering tasks involving complex datasets. Once the embeddings are reduced to a manageable number of dimensions, clustering algorithms such as k-means can be applied more effectively. K-means operates by partitioning the data into k clusters, minimizing the within-cluster variance based on a chosen distance metric, such as Euclidean or Manhattan distance. The choice of distance metric can influence the clustering outcome, and it is often selected based on the characteristics of the data and the specific objectives of the analysis [6]. In the context of accident narrative analysis, the combination of semantic-rich embeddings and dimensionality reduction enables k-means to uncover latent patterns and groupings that may correspond to distinct types of incidents, contributing factors, or outcomes [16][3]. The interpretability and utility of clustering results in text mining are further enhanced by integrating metadata variables, such as job experience, incident type, or location, with the textual embeddings [27]. This enrichment allows for a more comprehensive analysis, as clusters can be examined not only in terms of their textual content but also with respect to relevant contextual factors. For example, the importance of job experience in predicting injury outcomes has been highlighted, suggesting that combining such metadata with narrative embeddings can yield deeper insights into accident causation and prevention strategies [29]. Moreover, the application of clustering in text mining extends beyond mere pattern discovery. It facilitates the automatic generation of classification outcomes from similarity data, enabling the identification of key risk factors and the development of targeted interventions [3]. The use of unsupervised machine learning techniques, including clustering, has proven effective in analyzing large-scale accident datasets, revealing trends and associations that may not be apparent through manual analysis or traditional statistical methods [16][27]. Ganguli et al. [12] state that NLP-based clustering approaches have been instrumental in uncovering hazardous practices in industrial settings by analyzing narrative data collected from workers. Despite these advancements, challenges remain in evaluating the quality and validity of clustering results in text mining. Traditional metrics for assessing clustering success, such as accuracy or root-mean-square error, may be insufficient or misleading, particularly when dealing with complex, high-dimensional data [15][6]. Alternative validation strategies, including the examination of cluster coherence and the use of domain-specific knowledge, are often necessary to ensure the practical relevance of the identified clusters [15][3]. In summary, clustering approaches in text mining have been significantly enhanced by the integration of semantic-rich embeddings, dimensionality reduction techniques like UMAP, and robust clustering algorithms such as k-means. This framework enables the discovery of meaningful patterns and trends in accident narratives, providing valuable insights for safety analysis and risk management [6][17][16].

7.2 K-means Clustering: Fundamentals and Adaptation

K-means clustering is a widely adopted unsupervised learning algorithm that partitions data into k distinct clusters by minimizing the within-cluster sum of squares. The algorithm iteratively assigns each data point to the nearest cluster centroid and updates the centroids based on the mean of the assigned points. This process continues until convergence, typically when cluster assignments no longer change or the reduction in the objective function falls below a threshold. The fundamental objective of k-means is to ensure that intra-cluster similarity is maximized while inter-cluster similarity is minimized, which is particularly effective for datasets where clusters are approximately spherical and equally sized. In the context of accident narrative analysis, k-means clustering has been leveraged to uncover latent patterns and groupings within high-dimensional feature spaces. The application of k-means to such data, however, is not without challenges. High-dimensional embeddings, such as those generated by pre-trained SBERT models, often contain redundant or noisy information that can obscure meaningful cluster structures. To address this, dimensionality reduction techniques like UMAP are employed prior to clustering. UMAP projects the high-dimensional embeddings into a lower-dimensional space, preserving both local and global data structure, which enhances the performance and interpretability of k-means clustering. Yuta Hozumi et al. state that dimension reduction using UMAP or t-SNE before applying k-means clustering leads to improved clustering accuracy compared to using the original high-dimensional features. This improvement is attributed to the ability of UMAP to maintain the intrinsic geometry of the data, thereby facilitating the identification of coherent clusters. Furthermore, UMAP demonstrates stable performance across various reduction ratios, making it a reliable choice for preprocessing in clustering pipelines [6]. The stability and reliability of UMAP in reducing dimensions are particularly advantageous when dealing with large-scale accident datasets, where computational efficiency and cluster quality are critical. The adaptation of k-means clustering to accident analysis is further exemplified by its integration with semantic-rich embeddings. By combining sentence embeddings derived from SBERT with relevant metadata, the feature space becomes more informative, capturing both the contextual semantics of accident narratives and structured attributes from the MSHA dataset. This enriched representation allows k-means to form clusters that reflect nuanced accident patterns, which may not be apparent through traditional attribute-based clustering alone [11][23]. For instance, Fan Zhang [11] outlines a workflow where text preprocessing and word embedding generation are foundational steps before clustering, highlighting the importance of robust feature engineering in the clustering process. Moreover, the effectiveness of k-means clustering in accident pattern discovery is influenced by the quality of feature extraction and preprocessing. Nanyonga et al. [23] emphasize that transforming textual data into numerical features, such as through TF-IDF or Word2Vec, is essential for enabling machine learning models to operate effectively on narrative data.

The combination of these features with k-means clustering supports the identification of accident types, contributing factors, and emerging trends within the dataset. The literature also points to the necessity of careful parameter selection and validation in k-means clustering. The choice of k, the number of clusters, is often determined through methods such as the elbow method or silhouette analysis, ensuring that the resulting clusters are both meaningful and interpretable [6]. Additionally, the initialization of cluster centroids can impact the convergence and quality of the final clusters, prompting the use of strategies like k-means++ for improved initialization. In accident analysis, k-means clustering has been applied to various domains, including construction and mining safety, to categorize accident types, identify precursors, and support risk assessment [16][1]. Vaibhav Raj [16] describes the use of unsupervised machine learning, including k-means, for clustering accident attributes, which aids in the systematic analysis of safety incidents. Zeqiri Kemajl et al. [1] highlight that the accuracy of accident prediction and risk assessment is more dependent on the processing and correlation of data than on the specific clustering methodology employed, underscoring the importance of comprehensive data preparation and feature integration. In summary, the adaptation of k-means clustering to the analysis of accident narratives involves a multi-step process: generating semantic-rich embeddings, reducing dimensionality with UMAP, and applying k-means to uncover meaningful clusters. This approach leverages the strengths of each component, semantic representation, dimensionality reduction, and unsupervised clustering, to provide deeper insights into accident patterns and trends, surpassing the capabilities of traditional clustering methods [6][11][23].

7.3 Evaluation Metrics for Clustering Quality

Evaluating the quality of clustering is essential for validating the effectiveness of unsupervised learning approaches in accident pattern discovery. The assessment of clustering outcomes typically involves both internal and external metrics, each providing unique perspectives on the structure and interpretability of the resulting clusters. Internal evaluation metrics focus on the intrinsic properties of the clusters, such as compactness and separation. One widely used internal metric is the Within-Cluster Sum of Squares (WCSS), which quantifies the variance within each cluster. Lower WCSS values indicate that data points within a cluster are closely grouped, suggesting higher cluster cohesion. The elbow method leverages WCSS by plotting it against the number of clusters and identifying the inflection point, which is considered optimal for cluster selection [6]. This approach is particularly useful when applying k-means clustering, as it provides a visual and quantitative means to determine the most appropriate number of clusters for the dataset. Another important aspect of internal evaluation is the preservation of latent semantic relationships within the data after dimensionality reduction. The authors of indicate that reducing the dimensionality of embeddings, for example through UMAP, can enhance cluster quality while maintaining the underlying semantic structure. This is crucial when working with high-dimensional sentence embeddings, as it ensures that the reduced representations still capture meaningful relationships between accident narratives. External evaluation metrics, on the other hand, compare the clustering results to ground truth labels or external criteria. The Adjusted Rand Index (ARI) is a prominent example, measuring the similarity between the predicted clusters and a reference classification. Higher ARI values reflect better alignment with known categories. Radu et al. state that the ARI for k-means clustering improves significantly when using semantically rich embeddings such as Doc2Vec, compared to more traditional representations like TF-IDF. This suggests that the choice of embedding method directly impacts the interpretability and accuracy of clustering outcomes. Adjusted Mutual Information (AMI) is another external metric that quantifies the agreement between the clustering and a reference partition, adjusted for chance. The results in [10] demonstrate that the use of advanced embeddings and dimensionality reduction can substantially increase AMI, indicating more meaningful and robust clusters. In the context of semantic similarity, cosine similarity is often employed to assess the closeness of sentence embeddings within clusters. Reimers et al. [17] highlight that Spearman's rank correlation between cosine similarity scores and gold-standard semantic relatedness labels provides a more reliable measure than Pearson correlation, especially for sentence embeddings derived from models like SBERT or BERT. This metric is particularly relevant when the goal is to ensure that clusters group together narratives with similar semantic content. The interpretability of clusters is further enhanced by post-processing techniques, such as visualizing clusters in two-dimensional space using t-SNE or UMAP. Rose et al. outline that such visualizations facilitate the identification of overarching trends and relationships between clusters and metadata variables, supporting a more nuanced evaluation of clustering quality. Moreover, the integration of metadata variables with semantic embeddings allows for a richer evaluation framework. By examining the correspondence between clusters and metadata attributes, researchers can assess whether the discovered patterns align with known accident types or reveal novel insights [15]. This approach moves beyond purely statistical metrics, incorporating domain knowledge into the evaluation process. The choice of feature extraction methods also influences clustering quality. Nanyonga et al. [23] emphasize that preprocessing steps such as stopword removal and the use of advanced feature extraction techniques like TF-IDF or Word2Vec are critical for generating high-quality numerical representations, which in turn affect the coherence and distinctiveness of clusters. The effectiveness of dimensionality reduction techniques, such as PCA, UMAP, and t-SNE, can be quantitatively assessed by comparing clustering accuracy before and after their application. Hozumi et al. [6] provide empirical evidence that clustering accuracy improves when dimensionality reduction is applied prior to k-means clustering, underscoring the importance of this step in the analytical pipeline. Finally, the evaluation of clustering quality is not limited to numerical metrics. The practical utility of clusters in accident analysis is demonstrated by their ability to facilitate the automatic categorization of narratives and support metadata-based analyses [15]. This underscores the value of combining quantitative metrics with qualitative assessments to ensure that clustering results are both statistically sound and operationally meaningful.

8 Framework for Integrated Accident Pattern Mining

8.1 Workflow Overview

The workflow for integrated accident pattern mining begins with the extraction of accident narratives and associated metadata from the MSHA accident database. These narratives, which contain unstructured textual descriptions of incidents, are processed using pre-trained SBERT models to generate dense, semantic-rich sentence embeddings. SBERT, as a supervised extension of BERT, is specifically trained on sentence pairs and leverages Siamese and triplet network architectures to produce embeddings that capture nuanced semantic relationships between sentences [18]. This approach enables the transformation of complex accident narratives into high-dimensional vector representations that encode both syntactic and semantic information, facilitating downstream analysis. To enhance the analytical power of these embeddings, they are concatenated with structured metadata variables from the MSHA dataset. Metadata may include categorical and numerical features such as accident type, location, time, and worker demographics. The integration of these variables with the SBERT-derived embeddings results in a comprehensive feature set that encapsulates both the narrative context and the structured attributes of each accident case. This fusion of data types is crucial for capturing the multifaceted nature of mining accidents, as it allows for the simultaneous consideration of textual and non-textual factors in subsequent analyses [1][2]. Given the high dimensionality of the combined feature space, dimensionality reduction is a necessary step to facilitate effective clustering and visualization. Uniform Manifold Approximation and Projection (UMAP) is employed for this purpose, as it is well-suited for preserving both local and global structure in the data when projecting to lower dimensions. UMAP enables the transformation of the high-dimensional embeddings into a two- or three-dimensional space, making it possible to visually inspect the distribution of accident cases and to identify latent patterns that may not be apparent in the original feature space [11]. The use of dimensionality reduction also mitigates the curse of dimensionality, which can adversely affect the performance of clustering algorithms and the interpretability of results [10]. Following dimensionality reduction, k-means clustering is applied to the low-dimensional representations to partition the accident cases into distinct groups based on their semantic and metadata-driven similarities. The k-means algorithm assigns each case to the nearest cluster centroid, effectively grouping accidents with similar characteristics and narrative content. This clustering step is instrumental in uncovering recurrent patterns, trends, and outlier events within the accident data, providing actionable insights for safety management and risk mitigation [1][2]. The clusters can be further analyzed to identify common risk factors, accident types, or injury severities, and to inform targeted interventions. The workflow is iterative and data-driven, allowing for the refinement of each step based on empirical results and domain knowledge. Statistical analysis and visualization tools are integrated throughout the process to validate the quality of embeddings, the effectiveness of dimensionality reduction, and the coherence of clusters. This comprehensive approach leverages advances in natural language processing, machine learning, and statistical analysis to move beyond traditional, manual methods of accident investigation, offering a scalable and reproducible framework for mining accident pattern discovery [18][11][1][2].

8.2 Data Preprocessing and Cleaning

Data preprocessing and cleaning are essential steps to ensure the reliability and interpretability of accident pattern mining using integrated narrative and metadata sources. The initial phase involves the acquisition and consolidation of accident narratives, which are often unstructured and require systematic transformation into analyzable formats. This process typically starts with the collation of narrative texts from the MSHA accident database, where each narrative is associated with corresponding metadata variables such as date, location, and accident type [20][16]. The integration of these heterogeneous data sources is crucial for enriching the semantic context of each accident record, enabling more comprehensive downstream analysis [15]. A fundamental aspect of preprocessing is the removal of irrelevant or redundant information from the narrative texts. Stopword filtering is a widely adopted technique, where common words that do not contribute meaningful information to the analysis are excluded. For instance, the use of NLTK's stopword list allows for the systematic elimination of such terms, resulting in a more focused representation of the accident narratives [21]. This step is particularly important for reducing noise and enhancing the quality of the extracted features, which directly impacts the performance of subsequent embedding and clustering algorithms. In addition to stopword removal, further text normalization procedures are applied. These may include lowercasing, punctuation removal, and stemming or lemmatization, all of which serve to standardize the textual data and minimize variability arising from linguistic differences [16]. The cleaned narratives are then paired with structured metadata, forming a unified dataset that captures both the semantic richness of the narratives and the categorical or numerical attributes from the metadata [20]. Handling missing or incomplete data is another critical consideration. Accident reports may contain gaps in either the narrative or metadata fields. Strategies such as imputation, exclusion, or the use of indicator variables are employed to address these deficiencies, ensuring that the dataset remains robust for analysis. The authors of [4] indicate that rigorous validation using annotated datasets is essential for quantifying the accuracy of preprocessing steps and identifying potential sources of misclassification or data loss. Balancing the dataset is also necessary, especially when certain accident types or causes are underrepresented. Techniques such as oversampling can be applied during preprocessing to mitigate class imbalance, thereby improving the generalizability of the clustering and classification models. Zhang et al. [11] state that data balancing, in conjunction with architectural adjustments to the neural network, can significantly enhance the efficiency and effectiveness of the analysis. The integration of narrative and metadata information not only enriches the dataset but also addresses the limitations of relying solely on structured data. For example, narrative texts often contain details about safety equipment or contextual factors that are absent from tabular metadata, leading to more accurate classification and pattern discovery. Yedla et al. [29] highlight that models trained on narratives can capture subtle cues and contextual information, which are critical for precise accident categorization. Prior to embedding generation, the preprocessed and cleaned dataset is subjected to further quality checks to ensure consistency and completeness. This includes verifying the alignment between narrative and metadata fields, as well as ensuring that all records conform to the expected data schema. According to [15], the establishment of robust relationships between narratives and metadata is instrumental in uncovering nontrivial patterns and trends within the accident data. Finally, the prepared dataset is ready for the application of advanced natural language processing techniques, such as generating sentence embeddings using pre-trained SBERT models. The quality of these embeddings, and the insights derived from subsequent dimensionality reduction and clustering, are highly dependent on the rigor and thoroughness of the preprocessing and cleaning steps. Correlation analysis, as outlined in [22], can further enhance the evaluation of feature relationships, supporting more effective text mining and knowledge extraction from the integrated dataset.

8.3 Combining Metadata and Narrative Embeddings

Combining metadata with narrative embeddings represents a significant advancement in the analysis of mining accident data, as it enables the integration of structured and unstructured information to uncover complex patterns that may otherwise remain hidden. The use of pre-trained SBERT models to generate sentence embeddings from accident narratives allows for the capture of nuanced semantic information, reflecting the specific context and sequence of events described in each report. These embeddings, by virtue of their contextual richness, provide a dense representation of the textual data, which is essential for downstream analytical tasks such as clustering and pattern recognition [28]. However, narrative data alone may not fully encapsulate all relevant factors influencing accident outcomes. Metadata variables, such as time of incident, worker age, employment type, mine type, and accident type, offer critical structured information that can contextualize the narrative content. For instance, certain predictors like age and employment type have been shown to influence injury susceptibility, and their integration with narrative embeddings can enhance the interpretability and predictive power of analytical models [25]. The combination of these two data modalities, semantic-rich embeddings and structured metadata, enables a more holistic representation of each accident case. The process of combining these data types typically involves concatenating the high-dimensional narrative embeddings with appropriately encoded metadata features. This fusion creates a composite feature space that reflects both the semantic content of the narratives and the categorical or numerical attributes of the metadata. The challenge lies in ensuring that the integration does not introduce bias or redundancy, and that the resulting feature space remains amenable to dimensionality reduction and clustering. Dimensionality reduction techniques such as UMAP are particularly well-suited for this task, as they can project the high-dimensional composite embeddings into a lower-dimensional space while preserving both local and global structure. This step is crucial for facilitating effective clustering with algorithms like k-means, which rely on meaningful distance metrics in the reduced space. The integration of metadata and narrative embeddings also addresses limitations observed in traditional approaches that rely solely on structured data or manual feature engineering. For example, classical methods often struggle to identify the most informative features from unstructured text, especially when the narratives are highly variable and context-dependent [29]. By leveraging pre-trained language models, the approach bypasses the need for external linguistic tools or handcrafted features, resulting in a more scalable and portable solution across different datasets and languages [28]. Furthermore, the combined feature space supports unsupervised learning techniques, such as clustering, which can reveal latent groupings and trends in the accident data that may correspond to underlying causal factors or risk profiles [15]. The value of this integrated approach is further underscored by its ability to support data-driven safety analysis and risk assessment. By clustering accident cases based on both narrative content and metadata, it becomes possible to identify recurring patterns, high-risk scenarios, and potential gaps in safety protocols. This, in turn, can inform targeted interventions and policy recommendations aimed at preventing future incidents. The systematic framework established through this methodology aligns with the broader objective of mining risk assessment, which seeks to combine severity and probability metrics with detailed incident characterization [1]. Moreover, the visualization of metadata parameters in relation to narrative clusters provides actionable insights for safety practitioners and decision-makers [15]. In summary, the combination of metadata and narrative embeddings, facilitated by advanced natural language processing and dimensionality reduction techniques, offers a robust and innovative framework for mining accident pattern analysis. This integrated approach not only enhances the granularity and depth of accident investigations but also paves the way for more effective and proactive safety management strategies [28][29][1][25][15].

8.4 Dimensionality Reduction and Clustering Pipeline

Dimensionality reduction and clustering are essential components in the analysis of high-dimensional semantic embeddings derived from accident narratives. The integration of pre-trained SBERT models enables the extraction of dense, low-dimensional representations from textual data, which are further enriched by incorporating structured metadata variables. This process results in a high-dimensional feature space that, while information-rich, poses challenges for direct analysis and visualization due to the curse of dimensionality and computational inefficiency [18][33]. To address these challenges, dimensionality reduction techniques are employed to project the high-dimensional embeddings into a lower-dimensional space while preserving the intrinsic structure and relationships within the data. Among various methods, UMAP (Uniform Manifold Approximation and Projection) is particularly effective for this purpose, as it maintains both local and global data structure, facilitating the subsequent clustering analysis. The use of UMAP allows for the visualization of complex semantic relationships and the identification of latent patterns that may not be apparent in the original high-dimensional space [18][15]. Traditional dimensionality reduction approaches, such as Principal Component Analysis (PCA) and Random Projection (RP), have been widely used for text and image data. However, RP, for instance, is computationally less expensive and less susceptible to the curse of dimensionality compared to other methods, making it suitable for large-scale applications [33]. Despite these advantages, neural network-based embeddings, such as those produced by SBERT, offer a more nuanced capture of semantic features, resulting in more efficient and meaningful representations for downstream tasks [18][17]. The authors of [26] highlight that distributed word embeddings address issues of word order loss and high dimensionality inherent in traditional one-hot encoding, further supporting the adoption of advanced embedding techniques in this framework. Once the embeddings are reduced to a manageable dimensionality, clustering algorithms can be applied to uncover patterns and groupings within the accident data. K-means clustering is a widely adopted method due to its simplicity and effectiveness in partitioning data into distinct clusters based on similarity in the reduced feature space. The process of clustering benefits from the reduced overlap and increased distinction between clusters, as observed when the number of clusters is increased and the embeddings are well-separated in the low-dimensional space [2]. This separation is crucial for meaningful interpretation and subsequent analysis of accident patterns. Alternative clustering approaches, such as DBSCAN, can also be considered, especially when dealing with high-dimensional textual data. DBSCAN's flexibility in pairing with various distance functions, such as cosine dissimilarity, makes it particularly suitable for clustering semantic embeddings, as cosine distance is more appropriate for high-dimensional text data than Euclidean distance [10]. The choice of distance metric and clustering algorithm should be informed by the specific characteristics of the data and the analytical objectives. Visualization techniques, such as t-SNE and UMAP, play a significant role in interpreting the results of dimensionality reduction and clustering. These methods enable the projection of complex data structures into two or three dimensions, allowing for intuitive exploration of cluster boundaries and relationships among accident narratives [15]. Visual representations, including word clouds and cluster plots, provide at-a-glance insights into the thematic content and distribution of clusters, supporting the identification of key trends and recurring patterns in the accident data [13][15]. The integration of semantic-rich embeddings, advanced dimensionality reduction, and robust clustering methods constitutes an innovative pipeline for accident pattern mining. This approach surpasses traditional methods by capturing deeper semantic relationships and enabling the discovery of nuanced patterns that inform safety analysis and intervention strategies [18][17][26]. The combination of these techniques not only enhances computational efficiency but also provides a more comprehensive understanding of the underlying factors contributing to mining accidents.

8.5 Post-clustering Analysis of Accident Patterns

8.5.1 Interpretation of Cluster Characteristics

The interpretation of cluster characteristics following the application of UMAP-based dimensionality reduction and k-means clustering to SBERT-embedded accident narratives, enriched with metadata, is a critical step in extracting actionable knowledge from mining accident data. The integration of semantic-rich sentence embeddings with structured metadata variables enables the identification of nuanced patterns that are not readily apparent through conventional analysis techniques. After clustering, each group of accident narratives can be examined for internal coherence, thematic content, and correlation with metadata attributes, providing a multi-faceted understanding of accident typologies. Clusters derived from this integrated approach often exhibit varying degrees of internal homogeneity and metadata correlation. Some clusters may be well-defined, with clear thematic or causal coherence, while others remain diffuse or ambiguous, necessitating further sub-clustering or post-hoc analysis to elucidate their underlying structure. For instance, Rose et al. indicate that even after rigorous hierarchical clustering, certain sub-clusters persist as imprecise, lacking clear trends when visualized through metadata bar plots. This ambiguity highlights the importance of post-processing steps, such as correlation analysis, to uncover latent driving factors within these clusters. In some cases, large and indistinct clusters may contain smaller, more precise sub-clusters that exhibit strong correlations with specific metadata variables, such as accident type, location, or severity, as observed in the analysis of cluster 2, which, despite its size and lack of crisp definition, contains sub-structures with high metadata correlation [15]. The interpretability of clusters is further enhanced by examining the distribution of key terms and topics within each group. Nanyonga et al. outline that non-negative matrix factorization (NMF) can be used to decompose document-term matrices, yielding interpretable topic representations that align with the most frequent terms in each cluster. This approach facilitates the extraction of meaningful insights from accident narratives, as thematic clusters can be characterized by their associated keywords, which are often visualized using word clouds to provide an intuitive overview of cluster content [23][13]. Such thematic analysis is particularly valuable in the context of aviation and mining safety reports, where clusters may correspond to specific operational scenarios, hazard types, or procedural failures. The relationship between cluster membership and metadata variables is a key aspect of post-clustering interpretation. By correlating cluster assignments with structured attributes such as injury severity, root cause, or accident location, it becomes possible to identify clusters that are enriched for particular risk factors or outcomes. For example, Li et al. demonstrate that certain topics, while representing a small proportion of overall narratives, are disproportionately associated with severe outcomes, such as crashes involving vulnerable road users leading to major injuries [7]. This type of analysis enables targeted interventions by highlighting clusters that warrant further investigation or preventive action. In addition, the use of advanced dimensionality reduction techniques such as UMAP allows for the preservation of both local and global data structure in the low-dimensional embedding space, facilitating the identification of predominant clusters and their spatial relationships. Hozumi et al. show that applying UMAP to high-dimensional distance matrices, followed by k-means clustering, reveals distinct groupings that can be mapped to geographic or operational patterns, as evidenced by the identification of six predominant clusters in United States accident data [6]. The elbow method is often employed to determine the optimal number of clusters, ensuring that the resulting groupings are both statistically robust and interpretable. The interpretive process is not without challenges. Some clusters may remain ambiguous, with low correlation to available metadata or lacking clear thematic coherence. In such cases, additional post-processing, including the exploration of sub-clusters or the application of alternative clustering algorithms, may be necessary to refine the analysis [15]. Furthermore, the integration of unsupervised clustering with supervised classification results, as described by Hozumi et al., allows for the quantitative evaluation of cluster quality and the comparison of different dimensionality reduction and clustering strategies [6]. This iterative approach ensures that the final cluster interpretations are both data-driven and contextually meaningful. The interpretability of clusters is also enhanced by the use of network-based approaches, where the relationships among risk factors and accident outcomes are modeled as interconnected structures. Li et al. describe how network analysis can be used to split the structure into distinct factors, enabling the identification of relationships that lead to specific accident types [8]. This network perspective complements the cluster-based approach by providing a relational view of the factors underlying accident patterns. Finally, the comprehensive nature of the dataset, as exemplified by the extraction of accident reports spanning multiple decades, ensures that the identified clusters reflect the evolving landscape of workplace hazards and safety measures [4]. The integration of semantic, structural, and temporal dimensions in the interpretation of cluster characteristics provides a holistic framework for understanding and mitigating mining accident risks.

8.5.2 Trend Analysis Across Clusters

Trend analysis across clusters, following the application of k-means to UMAP-reduced SBERT embeddings enriched with MSHA metadata, enables the identification of nuanced accident patterns that are otherwise obscured in high-dimensional or unstructured data. The clustering process, by grouping semantically similar accident narratives, provides a foundation for systematic comparison of accident characteristics and their evolution over time or across operational contexts [6][8]. The interpretability of clusters is enhanced by the semantic richness of SBERT embeddings, which capture subtle linguistic cues and contextual information from accident narratives. This semantic depth, when combined with structured metadata, allows for the emergence of clusters that are not solely defined by surface-level features but by underlying thematic or causal similarities [11]. For instance, clusters may reveal recurring accident types, such as equipment-related incidents or falls, and their association with specific operational conditions or temporal trends. The integration of metadata variables, such as location, time, or equipment type, further refines the analysis, enabling the detection of patterns that span both narrative content and structured attributes [12][25]. The reduction of embedding dimensionality via UMAP is instrumental in preserving the global and local structure of the data, which is critical for meaningful clustering. Hozumi et al. [6] state that UMAP, when used prior to k-means, enhances clustering accuracy by maintaining the integrity of semantic relationships in a lower-dimensional space. This facilitates the visualization and interpretation of clusters, making it possible to track the prevalence of certain accident types or risk factors across different clusters and over time. Cluster analysis also supports the categorization of safety risk factors, as demonstrated by Li et al. [8]. By assigning accidents to clusters based on shared semantic and metadata features, it becomes feasible to quantify the distribution of risk factors and responsible participants within each cluster. This approach allows for the identification of clusters that are dominated by specific causal mechanisms or participant roles, providing actionable insights for targeted intervention strategies. The trend analysis is further enriched by examining the co-occurrence of thematic factors within and across clusters. Maa and Chen [3] highlight the value of counting co-occurrences to uncover associations between different types of information extracted from accident texts. This method reveals not only the dominant themes within clusters but also the interplay between various risk factors, contributing to a more holistic understanding of accident causation. The challenge of unstructured and imprecise language in accident narratives, as noted by Zhang [11], underscores the importance of robust semantic embedding techniques. The SBERT-based approach mitigates some of these challenges by capturing deeper semantic relationships, yet certain labels or accident types may still exhibit ambiguity or overlap across clusters. This necessitates careful interpretation of cluster boundaries and the potential for manual validation in cases where automated classification remains uncertain. The integration of clustering results with descriptive modeling, as outlined by Li et al. [8], enables each accident to be explained by the factors characterizing its assigned cluster. This not only aids in retrospective analysis but also supports predictive modeling and the development of early warning systems. The clear categorization of risk factors and accident types within clusters provides a structured basis for monitoring trends, evaluating the effectiveness of safety interventions, and informing policy decisions. In summary, the trend analysis across clusters derived from UMAP-reduced, SBERT-embedded accident narratives and metadata offers a comprehensive framework for uncovering latent patterns in mining accident data. This approach leverages advanced natural language processing, dimensionality reduction, and clustering techniques to move beyond traditional, surface-level analyses, enabling a deeper exploration of accident causation, risk factor distribution, and temporal or contextual trends [12][8][6][11][25].

8.5.3 Insights for Risk Management and Prevention

The integration of semantic-rich sentence embeddings with metadata variables, followed by dimensionality reduction and clustering, provides a nuanced perspective for risk management and prevention in mining accident analysis. By leveraging pre-trained SBERT models, the semantic content of accident narratives is captured in high-dimensional vector spaces, enabling the identification of subtle patterns that may not be apparent through traditional categorical or keyword-based approaches [17][11]. The subsequent application of UMAP for dimensionality reduction preserves the local and global structure of the data, facilitating the visualization and interpretation of complex relationships among accident cases [15]. Clustering with k-means on these reduced representations allows for the grouping of accidents with similar semantic and contextual characteristics, which is instrumental in uncovering recurrent risk factors and emerging trends [24]. The post-clustering analysis reveals that accidents sharing similar narrative structures and metadata often correspond to analogous risk scenarios, even when surface-level descriptors differ. This observation aligns with the findings that language used in accident reports can be highly overlapping across different risk categories, necessitating advanced semantic analysis to disentangle the underlying risk scenarios [34]. The ability to map clusters to specific risk profiles enables targeted risk mitigation strategies, as it becomes possible to associate particular patterns of incidents with operational, environmental, or human factors [3][1]. For instance, clusters characterized by frequent mentions of equipment malfunction, combined with metadata indicating certain mine types or shifts, can inform focused interventions in maintenance protocols or shift scheduling. The framework's capacity to synthesize narrative and structured data enhances the extraction of actionable knowledge from historical accident records. This is particularly valuable for risk management, as it supports the identification of leading indicators and precursors to severe incidents [9]. The clustering results can be cross-referenced with established risk analysis frameworks, such as Accimap, to validate and refine the categorization of risks and to ensure that the insights are grounded in both empirical data and theoretical models [3]. Moreover, the approach facilitates the automation of risk extraction, reducing reliance on subjective expert judgment, which is often limited by cognitive biases and incomplete exposure to the full spectrum of accident scenarios [2][9]. The interpretability of clusters is further enhanced by examining the most salient features, both semantic and metadata-driven, that define each group. This enables risk analysts to not only detect high-risk patterns but also to understand the contributing factors at a granular level. For example, the presence of certain keywords or phrases in conjunction with specific operational contexts can highlight systemic vulnerabilities that may otherwise go unnoticed [11][29]. The use of word embeddings and neural network models for classification has demonstrated effectiveness in matching safety measures to hazards, supporting the development of proactive hazard management systems. Additionally, the representativeness of the clustered accident types can be assessed by comparing the distribution of clusters with national or industry-wide accident statistics, ensuring that the insights are generalizable and not artifacts of sampling bias [3]. This validation step is crucial for translating analytical findings into practical risk management policies and prevention strategies. The clustering-based approach also supports the continuous improvement of risk management practices by enabling the monitoring of changes in accident patterns over time. As new data are incorporated, shifts in cluster composition or the emergence of novel clusters can signal evolving risks, prompting timely updates to safety protocols and training programs [15][1]. The adaptability of the framework to different data types and operational contexts further enhances its utility for dynamic risk assessment and prevention planning. In summary, the integration of semantic embeddings, dimensionality reduction, and clustering provides a robust foundation for extracting deep insights from accident data, supporting evidence-based risk management and the design of targeted prevention measures [2][17][15][34].

9 Comparative Analysis with Conventional Methods

9.1 Analysis Using TF-IDF, Word2Vec, and GloVe

Traditional text analysis methods such as TF-IDF, Word2Vec, and GloVe have been widely adopted for extracting features from accident narratives and related safety documents. Each of these approaches offers distinct mechanisms for representing textual data, which in turn influence their effectiveness in capturing semantic information and supporting downstream analytical tasks. TF-IDF, or term frequency-inverse document frequency, is a statistical measure that evaluates the importance of a word in a document relative to a corpus. It is particularly effective for identifying high-frequency terms that may serve as indicators of key factors or themes within accident reports. For instance, in the context of safety analysis, TF-IDF has been utilized to prioritize words from documents such as requests for information, with the mean TF-IDF value serving as a threshold to distinguish high-frequency terms. However, while TF-IDF is useful for highlighting salient vocabulary, its reliance on word frequency can limit its ability to capture deeper semantic relationships, especially when applied to diverse corpora or when the extraction aims require more nuanced understanding of context [31]. The method's performance is also sensitive to the specific characteristics of the dataset, necessitating adaptation and refinement for different applications. Word2Vec represents a significant advancement over purely frequency-based methods by learning distributed representations of words in a continuous vector space. Through training on large corpora, Word2Vec captures syntactic and semantic relationships between words, enabling the identification of patterns such as similarity and analogy. This capability is particularly valuable in accident analysis, where understanding the context and relationships between terms can reveal underlying causes and contributing factors. The effectiveness of Word2Vec, however, is influenced by the quality and quantity of the training data, as well as the preprocessing steps applied to the text. For example, the removal of stop-words and stemming are common preprocessing techniques that can enhance the quality of word embeddings by reducing noise and focusing on meaningful content [35]. Despite these strengths, Word2Vec operates primarily at the word level, which may limit its ability to capture the full semantic content of longer narratives or sentences. GloVe, or Global Vectors for Word Representation, extends the principles of Word2Vec by incorporating global word co-occurrence statistics from the entire corpus. This approach enables GloVe to generate word embeddings that reflect both local context and broader statistical patterns, potentially offering richer semantic representations. In comparative studies, GloVe and similar embedding models have been evaluated alongside other deep learning-based approaches, with their performance often benchmarked using metrics such as area under the curve (AUC) for classification tasks [11]. The dimensionality of the embeddings, as well as the aggregation functions applied across token representations, play a crucial role in determining the effectiveness of these models for specific analytical objectives [36]. For instance, the choice between unigram and bigram features, as well as the dimensionality of the embedding vectors, can significantly impact the ability to distinguish between different accident causes [11]. The application of these conventional methods in accident analysis is further shaped by the preprocessing and feature extraction strategies employed. The elimination of stop-words, stemming, and the use of word embeddings are standard practices that help to distill the most relevant information from raw narrative data [35]. Additionally, the integration of metadata and the use of advanced visualization techniques, such as those enabled by graph generation tools, can enhance the interpretability of the extracted features and support the identification of trends and hazardous scenarios [21]. The methodological parameters, including the definition and extraction of attributes from injury reports, also influence the overall effectiveness of the analysis [9]. While TF-IDF, Word2Vec, and GloVe each contribute valuable perspectives to the analysis of accident narratives, their limitations in capturing sentence-level semantics and integrating heterogeneous data sources have motivated the exploration of more advanced models. The comparative analysis of these methods provides a foundation for understanding the incremental benefits offered by newer approaches that leverage sentence embeddings and sophisticated clustering techniques. The authors of [36] indicate that the choice of embedding model and aggregation strategy is critical for achieving robust representations, especially when the goal is to move beyond surface-level patterns and uncover deeper insights into accident causation and prevention.

9.2 Advantages of SBERT-UMAP-K-means Integration

The integration of SBERT, UMAP, and k-means clustering presents several distinct advantages over conventional approaches for mining accident analysis. At the core of this framework lies the use of pre-trained SBERT models, which generate dense, semantically meaningful sentence embeddings from accident narratives. Unlike traditional bag-of-words or TF-IDF representations, SBERT embeddings capture contextual and syntactic nuances, enabling a more comprehensive encoding of narrative information. The pooling strategies available in SBERT, such as mean or max pooling, further enhance the flexibility and effectiveness of the resulting sentence vectors, allowing for adaptation to the specific characteristics of the dataset [17]. This semantic richness is particularly valuable when analyzing complex accident reports, where subtle linguistic cues may indicate underlying causes or contributing factors. By enriching these embeddings with metadata variables from the MSHA accident data, the approach leverages both unstructured textual information and structured contextual features. This fusion creates a high-dimensional feature space that encapsulates a broader spectrum of accident characteristics than text or metadata alone could provide [31]. However, such high-dimensional spaces can pose challenges for clustering and visualization due to the curse of dimensionality. Here, UMAP serves as a powerful dimensionality reduction technique, preserving both local and global data structure while projecting the embeddings into a lower-dimensional space suitable for clustering. UMAP has demonstrated superior performance in maintaining the integrity of data relationships compared to linear methods like PCA, especially when dealing with complex, nonlinear manifolds inherent in textual data. The ability of UMAP to retain meaningful neighborhood information ensures that clusters formed in the reduced space remain representative of the original semantic groupings. Once the embeddings are reduced, k-means clustering is applied to identify patterns and trends within the accident data. The effectiveness of k-means is significantly enhanced by the quality of the input features; when fed with UMAP-reduced SBERT embeddings, the algorithm can form more coherent and well-separated clusters, as evidenced by improved clustering accuracy and interpretability [6]. The use of cosine distance in variants such as spherical k-means further aligns with the nature of sentence embeddings, which are often normalized and benefit from angular similarity measures [10]. This synergy between embedding, reduction, and clustering components allows for the discovery of nuanced accident patterns that may be obscured by conventional methods relying solely on surface-level features or manual categorization. Compared to traditional approaches, which often depend on manual keyword extraction, basic statistical analysis, or shallow vectorization techniques, the SBERT-UMAP-k-means pipeline offers a more automated, scalable, and semantically informed framework. For instance, while methods like RAKE can extract important keywords from accident narratives, they lack the capacity to capture deeper semantic relationships or contextual dependencies between terms [30]. Similarly, conventional clustering based on TF-IDF or Doc2Vec representations may not achieve the same level of completeness or cluster quality as those based on SBERT embeddings, particularly when combined with advanced reduction techniques [10]. The integration of these modern methods thus enables a more granular and insightful analysis of accident data, facilitating the identification of latent trends, risk factors, and potential intervention points. Furthermore, the computational efficiency of the SBERT-UMAP-k-means pipeline is notable. SBERT models can be fine-tuned rapidly, often in less than 20 minutes, while still outperforming comparable sentence embedding methods [17]. UMAP, in turn, provides fast and scalable dimensionality reduction, making the approach suitable for large-scale mining datasets. The resulting clusters can be directly linked to actionable insights, supporting targeted safety interventions and risk mitigation strategies [1]. The authors of Xu et al. [31] indicate that integrating domain-specific lexicons and prioritizing high-frequency terms can further refine the analysis, ensuring that the most critical safety risk factors are highlighted. In summary, the SBERT-UMAP-k-means integration advances accident analysis by combining semantic depth, dimensionality reduction, and robust clustering. This framework not only enhances the interpretability and accuracy of pattern discovery but also streamlines the analytical workflow, offering a significant improvement over conventional methods in both methodological rigor and practical utility [1][6][31][17].

9.3 Limitations and Considerations

A critical examination of the proposed framework, which integrates pre-trained SBERT embeddings, metadata enrichment, UMAP-based dimensionality reduction, and k-means clustering, reveals several limitations and considerations that must be addressed when comparing it to conventional accident analysis methods. One of the primary challenges lies in the stability and interpretability of dimensionality reduction techniques. While UMAP demonstrates improved stability and clustering accuracy over methods such as t-SNE and PCA, its performance can still be sensitive to parameter choices and the intrinsic structure of the data. The reduced feature space, although more representative for clustering, may obscure certain nuances present in the original high-dimensional embeddings, potentially leading to the loss of subtle semantic relationships within accident narratives [6][33]. This limitation is particularly relevant when the goal is to capture complex, context-dependent patterns that may not be preserved after aggressive dimensionality reduction. The reliance on pre-trained SBERT models introduces another layer of complexity. Although these models are adept at capturing semantic information, their effectiveness is contingent upon the alignment between the pre-training corpus and the domain-specific language of mining accident reports. Domain adaptation remains a concern, as SBERT may not fully capture industry-specific terminology or context, which could affect the quality of the generated embeddings and, consequently, the downstream clustering results [2][20]. Furthermore, the process of combining sentence embeddings with structured metadata variables, while enriching the feature space, may introduce challenges related to feature scaling and integration, potentially impacting the interpretability and performance of subsequent analyses. Clustering with k-means, despite its widespread use and computational efficiency, is not without drawbacks. The algorithm assumes that clusters are spherical and equally sized, which may not hold true for the complex, heterogeneous patterns present in accident data. K-means is also sensitive to the initial selection of centroids and the choice of K, which can lead to suboptimal or unstable cluster assignments [23]. In the context of accident analysis, this may result in the formation of clusters that do not correspond to meaningful or actionable groupings, thereby limiting the practical utility of the findings. Another consideration is the interpretability of the resulting clusters. While the integration of semantic-rich embeddings and metadata can uncover latent patterns that are not easily accessible through traditional descriptive statistics, the high dimensionality and abstract nature of the embeddings can make it challenging to provide clear, human-interpretable explanations for why certain narratives are grouped together [2][20]. This opacity may hinder the translation of analytical insights into concrete safety interventions or policy recommendations. The framework's dependence on advanced natural language processing and machine learning techniques also necessitates significant computational resources and technical expertise. This requirement may pose barriers to adoption in organizations with limited access to such resources or expertise, especially when compared to more straightforward conventional methods such as logistic regression or decision trees, which are easier to implement and interpret [29][25]. Decision trees, for instance, offer transparency and speed but are prone to overfitting and may not capture the full complexity of the data, whereas logistic regression provides trend identification but lacks the depth needed for comprehensive accident analysis [29][25]. Moreover, the integration of narrative and structured data, while offering a more holistic view, introduces challenges related to data quality and consistency. Accident narratives are often unstructured, variable in length, and subject to reporting biases or inconsistencies, which can affect the reliability of the embeddings and the overall analysis. Human oversight remains essential to ensure that machine-generated insights are contextually valid and actionable, as automated methods may overlook subtle but critical factors embedded in the text [34]. Finally, the identification of clusters and sub-clusters with low correlation to available metadata highlights the limitations of relying solely on existing structured variables for interpretation.

Some clusters may be driven by latent factors not captured in the metadata, necessitating further exploration and possibly the inclusion of additional contextual information to fully understand the underlying causes of accidents [15]. This underscores the need for iterative refinement and validation of the analytical framework to ensure that it remains responsive to the evolving nature of accident data and organizational needs. In summary, while the proposed approach offers significant advancements over conventional methods by leveraging semantic-rich embeddings and advanced clustering, it is accompanied by limitations related to dimensionality reduction, model interpretability, domain adaptation, computational demands, and data quality. Addressing these considerations is essential for maximizing the framework's effectiveness and ensuring its practical relevance in mining accident analysis [6][29][2][23][20][34][25][33].

10 Broader Applications and Extensions

10.1 Applicability to Other Domains of Safety and Risk Analysis

The integration of pre-trained SBERT models for generating sentence embeddings, combined with metadata enrichment and advanced dimensionality reduction techniques such as UMAP, presents a versatile analytical framework that extends well beyond the mining sector. This methodology, which leverages semantic-rich representations and clustering algorithms like k-means, is inherently adaptable to a wide spectrum of safety and risk analysis domains. In the context of construction safety, the challenges associated with mapping large volumes of unstructured accident data into structured accident models are well recognized. The use of natural language processing (NLP) and clustering methods has already demonstrated potential for extracting essential information from narrative reports and expressing it within established accident causation frameworks. This approach is not limited to a single industry; it can be migrated to analyze systemic accidents in other regions or sectors, provided that the underlying data can be effectively clustered and interpreted using NLP-driven techniques [3]. The ability to cluster accident information using sentence embeddings allows for the identification of latent patterns and risk factors that may not be immediately apparent through traditional qualitative or quantitative analyses. Moreover, the application of advanced embedding models, such as those based on transformer architectures, has shown significant promise in automating the categorization of incident narratives and in identifying contributing factors in domains like highway construction safety. Smetana et al. highlight that state-of-the-art embedding models, including OpenAI's Ada and SBERT, have achieved top performance in clustering safety-related incidents, which underscores their utility in extracting actionable insights from textual data. The automation of incident categorization and the identification of contributing factors are critical for data-driven decision making, which is essential for accident prevention and safety improvement in various industries [2]. The generalizability of this framework is further supported by the fact that clustering analysis, when applied to safety risk factors extracted from accident reports, enables the classification of these factors into direct causes and management deficiencies. This process is not unique to mining but is applicable wherever narrative data on incidents or near-misses are available. By grouping similar risk factors, organizations can prioritize interventions and tailor safety programs to address the most pressing hazards [8]. The clustering of textual data, facilitated by robust preprocessing and normalization techniques, ensures that language inflections and variations do not obscure underlying patterns, making the approach suitable for multilingual and cross-domain applications [10]. However, the successful transfer of this methodology to other domains requires careful consideration of domain-specific variables and the expansion of existing databases to support more sophisticated NLP tasks. For instance, in construction safety research, there is a recognized need to refine databases and conduct systematic analyses that employ different accident causation models, thereby enriching the perspectives available for risk analysis [3]. The absence of comprehensive follow-up data and the reliance on limited methodological tools have been identified as significant weaknesses in traditional safety research, further emphasizing the value of data-driven, NLP-based approaches [32]. The adaptability of the proposed framework is also evident in its capacity to handle both structured and unstructured data. For example, binary features derived from categorical metadata can be seamlessly integrated with unstructured narrative embeddings, enhancing the predictive performance of safety incident models. This hybrid approach has been shown to improve the identification of key contributors to accidents, such as specific actions or environmental factors, which can be critical for targeted safety interventions [24]. Furthermore, the use of dimensionality reduction techniques like UMAP and t-SNE is not confined to mining or construction. These methods are widely applicable for visualizing and interpreting high-dimensional embeddings in any domain where large-scale textual or mixed-type data are present. Ayesha et al. [33] outline that t-SNE, for instance, is effective in reducing high-dimensional data to lower dimensions while preserving the structure necessary for meaningful clustering, which is essential for uncovering hidden trends in safety data. The broader applicability of this framework is also supported by advances in text analytics and the increasing availability of pre-trained language models. The deep bidirectionality and contextual understanding provided by models such as BERT and its derivatives enable more accurate extraction of semantic information from complex narratives, which is crucial for nuanced safety and risk analysis across diverse sectors [5]. The continuous evolution of embedding models and clustering algorithms will likely further enhance the transferability and effectiveness of this approach in new domains. In summary, the integration of semantic-rich embeddings, dimensionality reduction, and clustering analysis constitutes a robust and flexible methodology for safety and risk analysis. Its applicability extends to any domain where incident narratives and metadata are available, including but not limited to construction, transportation, manufacturing, and healthcare. The framework's ability to uncover deep patterns, automate categorization, and support data-driven decision making positions it as a valuable tool for advancing safety research and practice across multiple industries [33][32][24][3][10][2][5][8].

10.2 Potential for Real-time Monitoring and Decision Support

The integration of pre-trained SBERT models for generating sentence embeddings from accident narratives, when combined with metadata from MSHA accident data, presents significant opportunities for real-time monitoring and decision support in mining safety management. The semantic richness of SBERT embeddings enables the capture of nuanced contextual information from unstructured narrative data, which is often overlooked by traditional shallow models that primarily rely on word-level representations and bag-of-words approaches [18][17]. By leveraging these deep contextual embeddings, the system can more accurately represent the underlying causes and circumstances of accidents, facilitating the identification of subtle patterns and precursors that may signal emerging risks [37][22]. The dimensionality reduction step, implemented via UMAP, is crucial for transforming high-dimensional semantic representations into a form suitable for rapid clustering and visualization. This reduction not only enhances computational efficiency but also preserves the local and global structure of the data, which is essential for real-time applications where timely insights are required [17]. The subsequent application of k-means clustering allows for the dynamic grouping of similar accident scenarios, enabling the detection of clusters that may correspond to specific types of hazards, operational failures, or environmental conditions [3]. Such clustering, when performed on continuously updated data streams, can support the early identification of anomalous patterns or the escalation of incident types, providing actionable intelligence to safety managers and decision-makers [15]. The potential for real-time monitoring is further amplified by the ability of the framework to integrate both narrative and structured metadata. Metadata variables, such as location, time, equipment involved, and severity, enrich the semantic embeddings and allow for multi-faceted analysis. This integration supports the development of dashboards and alerting systems that can flag high-risk situations as they evolve, rather than relying solely on post-hoc analysis [7]. For instance, clusters associated with high-frequency accident types, such as explosions, fires, or poisonings, can be continuously monitored for changes in their composition or frequency, enabling proactive interventions [22]. Moreover, the framework's adaptability to new data and topics is supported by the generalization capabilities of SBERT and similar transformer-based models, which have demonstrated robust performance across diverse domains and tasks [5][17]. This adaptability is essential for real-time systems, as it allows the monitoring platform to remain effective even as the nature of mining operations or reporting practices evolve. The clustering and anomaly detection components can be recalibrated as new patterns emerge, ensuring that decision support remains relevant and responsive to current operational realities [15][3]. The descriptive and predictive analytics enabled by this approach can be harnessed not only for immediate response but also for strategic planning. For example, the identification of systemic challenges and human factors underlying accident clusters can inform targeted training, policy adjustments, and resource allocation [32][15]. The ability to synthesize narrative and quantitative data provides a comprehensive view of safety performance, supporting both short-term interventions and long-term improvements. In summary, the integration of semantic-rich embeddings, dimensionality reduction, and clustering within a unified framework offers a transformative approach to real-time monitoring and decision support in mining accident analysis. This methodology enables the continuous extraction of actionable insights from complex, heterogeneous data sources, supporting both operational safety and strategic risk management [22][15][17][37][18][7][32][3][5].

10.3 Future Directions in Semantic Accident Analysis

Future directions in semantic accident analysis are shaped by the rapid evolution of natural language processing, the increasing availability of large-scale accident datasets, and the growing demand for actionable insights in occupational safety. The integration of pre-trained SBERT models for sentence embeddings, as demonstrated in this research, opens several promising avenues for further exploration and refinement. One significant direction is the expansion of semantic analysis frameworks to incorporate multimodal data sources. While current approaches primarily focus on textual narratives and structured metadata, future work could integrate sensor data, images, or audio recordings from accident scenes. Such multimodal fusion has the potential to uncover latent patterns that are not accessible through text alone, thereby enhancing the granularity and reliability of accident causation models [15][22]. The authors of [15] indicate that combining different data modalities can reveal nontrivial patterns, suggesting that future research should prioritize the development of robust data fusion techniques tailored to the unique characteristics of mining and industrial safety datasets. Another promising trajectory involves the refinement of dimensionality reduction and clustering techniques. Although UMAP and k-means have proven effective for visualizing and grouping high-dimensional embeddings, there is scope for exploring alternative algorithms that may better capture the complex, nonlinear relationships inherent in accident data. For instance, the use of advanced manifold learning or graph-based clustering methods could provide more nuanced groupings, especially when dealing with heterogeneous data sources [22][33]. Sifeng Jinga et al. [22] state that correlation analysis and text clustering are essential for extracting latent knowledge, and future research could benefit from integrating these methods with semantic embeddings to improve the interpretability and actionability of discovered clusters. Interpretability remains a central challenge in semantic accident analysis. As models become more sophisticated, ensuring that their outputs are transparent and actionable for safety professionals is critical. Non-negative matrix factorization (NMF) has been highlighted for its interpretability in topic modeling of accident reports [13]. Future research could explore hybrid approaches that combine the semantic richness of transformer-based embeddings with the clarity of matrix factorization or rule-based extraction methods [21][13]. This would facilitate the translation of complex model outputs into practical recommendations for accident prevention and safety training. Automated keyword and hazard extraction from accident narratives is another area ripe for advancement. While rule-based systems have demonstrated effectiveness in identifying hazards without extensive training data [21], the integration of semantic embeddings could enable more context-aware extraction, capturing subtle or emerging risks that may not be covered by predefined term lists. The RAKE algorithm, for example, has shown promise in extracting key factors from traffic crash narratives [30], and its adaptation to mining accident data could enhance the detection of both direct and indirect risk factors. The reliability and generalizability of semantic models are also critical considerations for future research. Discrepancies between manual and automated coding, as noted in, highlight the need for rigorous validation protocols and the inclusion of diverse accident types and severity levels in training datasets. Expanding the scope of analysis to encompass a broader range of incidents, including near-misses and low-severity events, could improve the robustness of predictive models and support more comprehensive safety interventions [9]. Furthermore, the application of semantic analysis to real-time accident monitoring and early warning systems represents a transformative opportunity. By leveraging streaming data and continuously updated embeddings, it may become feasible to detect emerging safety trends and intervene proactively. This aligns with the broader goal of accident causation modeling, which seeks to identify and mitigate risks before they result in harm [3]. Finally, future research should address the ethical and practical implications of deploying advanced semantic analysis tools in operational settings. Ensuring data privacy, minimizing algorithmic bias, and providing clear guidance for end-users are essential for the responsible adoption of these technologies. The empathetic approach advocated in underscores the importance of centering worker safety and lived experience in the design and evaluation of analytical frameworks. In summary, the future of semantic accident analysis lies in the integration of multimodal data, the advancement of interpretable and robust modeling techniques, the automation of context-aware hazard extraction, and the translation of analytical insights into effective safety interventions. These directions promise to bridge existing gaps in accident research and practice, ultimately contributing to safer workplaces and more resilient safety cultures [32][30][15][22][3][21][33][13].

11 Conclusion

The integration of advanced natural language processing techniques, particularly the use of pre-trained Sentence-BERT models, with structured metadata and sophisticated dimensionality reduction methods such as UMAP, marks a significant advancement in the analysis of mining accident narratives. This comprehensive framework enables the transformation of unstructured textual data into semantically rich, high-dimensional embeddings that capture nuanced contextual and syntactic information. When combined with relevant metadata, these embeddings provide a holistic representation of accident cases, facilitating a deeper understanding of the multifaceted factors contributing to mining incidents.

Addressing the challenges posed by the high dimensionality of semantic embeddings, the application of UMAP effectively preserves both local and global data structures while reducing computational complexity. This dimensionality reduction is critical for enabling robust clustering through algorithms like k-means, which can then uncover latent patterns, thematic groupings, and emergent trends within the accident data. The resulting clusters not only reflect coherent accident typologies but also reveal associations with metadata variables such as injury severity, accident type, and operational context, thereby supporting targeted risk management and prevention strategies.

Comparative analyses demonstrate that this integrated approach surpasses traditional methods based on TF-IDF, Word2Vec, or GloVe by offering superior semantic representation, computational efficiency, and interpretability. The synergy between SBERT embeddings, UMAP reduction, and k-means clustering facilitates scalable and automated accident pattern discovery, reducing reliance on manual coding and subjective judgment. Nonetheless, considerations remain regarding the stability of dimensionality reduction, domain adaptation of embedding models, and the interpretability of complex clusters. Addressing these limitations through iterative refinement, domain-specific tuning, and enhanced visualization techniques will be essential for maximizing the practical utility of the framework.

Beyond mining, the methodology exhibits broad applicability across various safety-critical industries, including construction, transportation, and manufacturing, where unstructured narrative data and structured metadata coexist. Its adaptability to diverse datasets and operational contexts positions it as a valuable tool for advancing safety research and practice. Furthermore, the potential for real-time monitoring and decision support systems emerges from the framework's capacity to process continuous data streams, detect evolving risk patterns, and provide actionable insights to safety managers.

Looking forward, future developments may involve the incorporation of multimodal data sources, such as sensor readings and imagery, to enrich semantic analysis and enhance predictive capabilities. The exploration of alternative dimensionality reduction and clustering algorithms, alongside efforts to improve model interpretability and automate hazard extraction, will further strengthen the analytical pipeline. Ethical considerations, including data privacy and algorithmic transparency, must also guide the deployment of these technologies in operational environments.

Ultimately, this integrated approach represents a transformative step toward more effective, data-driven accident analysis and prevention. By harnessing the power of semantic embeddings, dimensionality reduction, and clustering, it enables a comprehensive and nuanced understanding of accident causation, supports evidence-based safety interventions, and contributes to the ongoing improvement of occupational health and safety outcomes in mining and beyond.

References

[1] Z. Kemajl, M. Stojance, I. Gzim, and M. L. Ledi, “Comprehensive analysis of the mining accident forecasting and risk assessment methodologies: Case study – stanterg mine”, *Mining of Mineral Deposits*, vol. 18, no. 2, pp. 11–17, Jun. 2024. DOI: 10.33271/mining18.02.011. [Online]. Available: https://doi.org/10.33271/mining18.02.011.

[2] M. Smetana, L. S. de Salles, I. Sukharev, and L. Khazanovich, “Highway construction safety analysis using large language models”, *Applied Sciences*, vol. 14, p. 1352, Feb. 2024. DOI: 10.3390/app14041352. [Online]. Available: https://doi.org/10.3390/app14041352.

[3] Z. Maa and Z.-S. Chen, “Mining construction accident reports via unsupervised nlp and accimap for systemic risk analysis”, *Automation in Construction*, vol. 161, p. 105 343, Mar. 2024. DOI: 10.1016/j.autcon.2024.105343. [Online]. Available: https://doi.org/10.1016/j.autcon.2024.105343.

[4] E. Ahmadi, S. Muley, and C. Wang, “Automatic construction accident report analysis using large language models (llms)”, *J Intell Constr*, Apr. 2025. DOI: 10.26599/JIC.2024.9180039. [Online]. Available: https://doi.org/10.26599/JIC.2024.9180039.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, Jun. 2019, pp. 4171–4186.

[6] Y. Hozumi, R. Wang, C. Yin, and G.-W. Wei, “Umap-assisted k-means clustering of large-scale sars-cov-2 mutation datasets”, *Computers in Biology and Medicine*, vol. 131, p. 104 264, Feb. 2021. DOI: 10.1016/j.compbiomed.2021.104264. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.104264.

[7] P. Li, S. Chen, L. Yue, Y. Xu, and D. A. Noyce, “Analyzing relationships between latent topics in autonomous vehicle crash narratives and crash severity using natural language processing techniques and explainable xgboost”, *Accident Analysis and Prevention*, vol. 203, p. 107 605, May 2024. DOI: 10.1016/j.aap.2024.107605. [Online]. Available: https://doi.org/10.1016/j.aap.2024.107605.

[8] J. Li, J. Wang, N. Xu, Y. Hu, and C. Cui, “Importance degree research of safety risk management processes of urban rail transit based on text mining method”, *Information*, Jan. 2018. DOI: 10.3390/info9020026.

[9] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, “Application of machine learning to construction injury prediction”, *Automation in Construction*, pp. 102–114, Jun. 2016. DOI: 10.1016/j.autcon.2016.05.016. [Online]. Available: https://doi.org/10.1016/j.autcon.2016.05.016.

[10] R.-G. Radu, I.-M. Rădulescu, C.-O. Truică, E.-S. Apostol, and M. Mocanu, “Clustering documents using the document to vector model for dimensionality reduction”, *IEEE*, Jun. 2020.

[11] F. Zhang, “A hybrid structured deep neural network with word2vec for construction accident causes classification”, *International Journal of Construction Management*, vol. 22, no. 6, pp. 1120–1140, Nov. 2019. DOI: 10.1080/15623599.2019.1683692. [Online]. Available: https://doi.org/10.1080/15623599.2019.1683692.

[12] R. Ganguli, P. Miller, and R. Pothina, “Effectiveness of natural language processing based machine learning in analyzing incident narratives at a mine”, *Minerals*, p. 776, Jul. 2021. DOI: 10.3390/min11070776. [Online]. Available: https://doi.org/10.3390/min11070776.

[13] A. Nanyonga, K. Joiner, H. Wasswa, G. Wild, and U. Turhan, *Comparative analysis of topic modeling techniques on atsb text narratives using natural language processing*, Mar. 2024.

[14] S. Onder, “Evaluation of occupational injuries with lost days among opencast coal mine workers through logistic regression models”, *Safety Science*, pp. 86–92, May 2013. DOI: 10.1016/j.ssci.2013.05.002.

[15] R. L. Rose, T. G. Puranik, and D. N. Mavris, “Natural language processing based method for clustering and analysis of aviation safety narratives”, *Aerospace*, Sep. 2020.

[16] V. R. aj, *Advanced application of text analytics in msha metal and nonmetal fatality reports*, Feb. 2020. [Online]. Available: https://www.researchgate.net/publication/349443633.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", Aug. 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[18] K. Babić, S. Martinčić-Ipšić, and A. Meštrović, "Survey of neural text representation models", *Information*, Oct. 2020. DOI: 10.3390/info11110511.

[19] B. Jiang and K. Wang, "Railway accident causation prediction with improved transformer model based on lexical information and contextual relationships", *Knowledge-Based Systems*, p. 111 897, May 2024. DOI: 10.1016/j.knosys.2024.111897. [Online]. Available: https://doi.org/10.1016/j.knosys.2024.111897.

[20] B. Zhong, X. Pan, P. E. Love, L. Ding, and W. Fang, "Deep learning and network analysis: Classifying and visualizing accident narratives in construction", *Automation in Construction*, vol. 113, p. 103 089, Feb. 2020. DOI: 10.1016/j.autcon.2020.103089. [Online]. Available: https://doi.org/10.1016/j.autcon.2020.103089.

[21] S. Ballal, K. A. Patel, and D. A. Patel, "Enhancing construction site safety: Natural language processing for hazards identification and prevention", *Journal of Engineering, Project, and Production Management*, pp. 00–14, Nov. 2024. DOI: 10.32738/JEPPM-2024-0014.

[22] S. Jinga, X. Liu, X. Gong, *et al.*, "Correlation analysis and text classification of chemical accident cases based on word embedding", *Process Safety and Environmental Protection*, pp. 698–710, Dec. 2021. DOI: 10.1016/j.psep.2021.12.038. [Online]. Available: https://doi.org/10.1016/j.psep.2021.12.038.

[23] A. Nanyonga, K. Joiner, H. Wasswa, G. Wild, and U. Turhan, *Exploring aviation incident narratives using topic modeling and clustering techniques*, Apr. 2024.

[24] R. Bridgelall and D. D. Tolliver, "Railroad accident analysis by machine learning and natural language processing", *Journal of Rail Transport Planning & Management*, vol. 29, p. 100 429, Dec. 2024. DOI: 10.1016/j.jrtpm.2023.100429. [Online]. Available: https://doi.org/10.1016/j.jrtpm.2023.100429.

[25] R. Amoako, J. Buaba, and A. Brickey, "Identifying risk factors from msha accidents and injury data using logistic regression", *Mining, Metallurgy & Exploration*, Oct. 2020. DOI: 10.1007/s42461-020-00347-x. [Online]. Available: https://doi.org/10.1007/s42461-020-00347-x.

[26] B. Zhong, X. Pan, P. E. Love, J. Sun, and C. Tao, "Hazard analysis: A deep learning and text mining framework for accident prevention", *Advanced Engineering Informatics*, vol. 46, Aug. 2020. DOI: 10.1016/j.aei.2020.101152. [Online]. Available: https://doi.org/10.1016/j.aei.2020.101152.

[27] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized machine learning techniques for prediction of occupational accidents", *Computers and Operations Research*, vol. 106, pp. 210–224, Mar. 2019. DOI: 10.1016/j.cor.2018.02.021. [Online]. Available: https://doi.org/10.1016/j.cor.2018.02.021.

[28] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bert's core : Evaluating text generation with bert", Feb. 2020. [Online]. Available: https://arxiv.org/abs/1904.09675.

[29] A. Yedla, F. D. Kakhki, and A. Jannesari, "Predictive modeling for occupational safety outcomes and days away from work analysis in mining operations", *International Journal of Environmental Research and Public Health*, Sep. 2020. DOI: 10.3390/ijerph17197054.

[30] S. Jaradat, A. Hossain, T. I. Alhadidi, M. Elhenawy, and H. I. Ashqar, *Exploring traffic crash narratives in jordan using text mining analytics*, X 20XX.

[31] N. Xu, L. Ma, Q. Liu, L. Wang, and Y. Deng, "An improved text mining approach to extract safety risk factors from construction accident reports", *Safety Science*, vol. 138, p. 105 216, Feb. 2021. DOI: 10.1016/j.ssci.2021.105216. [Online]. Available: https://doi.org/10.1016/j.ssci.2021.105216.

[32] A. D. Rafindadi, B. Kado, A. M. Gora, *et al.*, "Caught-in/between accidents in the construction industry: A systematic review", *Safety*, vol. 11, p. 12, Feb. 2025. DOI: 10.3390/safety11010012. [Online]. Available: https://doi.org/10.3390/safety11010012.

[33] S. Ayesha, M. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data”, *Information Fusion*, vol. 59, pp. 44–58, Jan. 2020. DOI: 10.1016/j.inffus.2020.01.005. [Online]. Available: https://doi.org/10.1016/j.inffus.2020.01.005.

[34] M. Figueres-Esteban, P. Hughes, and C. van Gulijk, “Visual analytics for text-based railway incident reports”, Jun. 2016. DOI: 10.1016/j.ssci.2016.05.009. [Online]. Available: http://dx.doi.org/10.1016/j.ssci.2016.05.009.

[35] P. Srinivasan, V. Nagarajan, and S. Mahadevan, “Mining and classifying aviation accident reports”, Jun. 2019. DOI: 10.2514/6.2019-2938. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2019-2938.

[36] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, “On measuring social biases in sentence encoders”, Mar. 2019. [Online]. Available: https://arxiv.org/abs/1903.10561.

[37] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, “Understanding the behaviors of bert in ranking”, *arXiv preprint arXiv:1904.07531v4*, Apr. 2019.