From Latent Topics to Causal Structures in Occupational Safety Data

Master of Science Thesis
University of Turku
Department of Computing
Machine Learning
2026
Santtu Hietamäki

UNIVERSITY OF TURKU
Department of Computing

Santtu Hietamäki: From Latent Topics to Causal Structures in Occupational Safety Data

Master of Science Thesis, 56 p.
Machine Learning
May 2026

Large amounts of occupational safety data involving natural language are collected every day. These data are often underutilized and preventive accident analysis is usually limited to more traditional statistical methods. The goal of this research is to find applications for the data using natural language processing methods. The thesis approaches this by asking whether incident topics can be identified from these data and whether these topics can be used for causal discovery. This study used occupational safety data from Sofor's TAVA safety system and proposed a potential method for causal discovery. Written incident reports were grouped with topic modeling using BERTopic. Aggregated time series for each topic were formed by counting incidents for each week. The time series were then compared with each other to identify potentially causally related topics. This was done using Granger causality and the PCMCI algorithm. Using LLM evaluation, the incidents were correctly grouped into topics around 60% of the time. From these incident topics, several potential causal links were found with PCMCI. These results suggest that it is feasible to develop a preliminary model for identifying patterns related to possible incident causes.

Keywords: occupational safety, incident reports, topic modeling, BERTopic, causal discovery, PCMCI

TURUN YLIOPISTO
Tietotekniikan laitos

Santtu Hietamäki: From Latent Topics to Causal Structures in Occupational Safety Data

Pro gradu -tutkielma, 56 s.
Machine Learning
Toukokuu 2026

Suuria määriä luonnollista kieltä sisältävää työturvallisuusdataa kerätään päivittäin. Tätä dataa hyödynnetään usein puutteellisesti, ja ennaltaehkäisevä onnettomuusanalyysi rajoittuu tavallisesti perinteisempiin tilastollisiin menetelmiin. Tämän tutkimuksen tavoitteena on löytää datalle uusia käyttötapoja luonnollisen kielen käsittelyn menetelmien avulla. Tutkielmassa tätä lähestytään kysymällä, voidaanko tästä datasta tunnistaa työturvallisuusaiheita ja voidaanko näitä aiheita käyttää kausaalisuhteiden tunnistamiseen. Tutkimuksessa hyödynnettiin Soforin TAVA-turvallisuusjärjestelmän työturvallisuusdataa ja ehdotettiin menetelmää kausaalianalyysiin. Kirjalliset työturvallisuustapahtumaraportit ryhmiteltiin aiheisiin BERTopic-topiikkimallinnuksen avulla. Jokaiselle aiheelle muodostettiin aggregoitu aikasarja laskemalla tapahtumien määrä viikoittain. Näitä aikasarjoja verrattiin tämän jälkeen keskenään, jotta voitiin tunnistaa mahdollisesti kausaalisesti toisiinsa liittyviä aiheita. Tämä tehtiin Granger-kausaalisuuden ja PCMCI-algoritmin avulla. LLM-arvioinnin perusteella tapahtumat ryhmiteltiin oikeisiin aiheisiin noin 60 % ajasta. Näistä tapahtuma-aiheista PCMCI löysi useita mahdollisia kausaalisia yhteyksiä. Tulokset viittaavat siihen, että on mahdollista kehittää alustava malli onnettomuuksien mahdollisiin syihin liittyvien rakenteiden tunnistamiseen.

Asiasanat: työturvallisuus, turvallisuusraportit, topiikkimallinnus, BERTopic, kausaalianalyysi, PCMCI

Contents

1 Introduction

1.1 Background

In 2022 there were 2.97 million non-fatal workplace accidents in the 27 member states of the European Union (causing at least four days of absence from work) and over 3200 fatal accidents in the same year (Eurostat, 2024). In Finland specifically there were 35 743 non-fatal accidents and 27 fatal accidents. When the accident rates are compared between EU countries, taking into account the size of the workforce and the level of economic activity, EU total was 1551 and Finland had similarly 1563 non-fatal incidents. The highest rates of fatal accidents were recorded in the construction (22.9%), transportation and storage (15.6%) and manufacturing (15.2%) sectors. Manufacturing had the highest rate of non-fatal accidents (18.0%). The European Union has made significant progress in improving workplace safety in recent decades, but there is still room for improvement. Companies are not only mandated by workplace safety regulations to prevent and minimize workplace accidents, accidents are also costly for the companies.

To address safety issues, several different data-driven methods have been proposed during the 20th century. Heinrich's Accident triangle was introduced in 1931, based on the idea that reducing minor accidents at the bottom of the triangle also reduces serious accidents at the top (Heinrich, 1941). In Taiichi Ohno's 5 Whys, implemented internally at Toyota through the 1950s, the goal is to iteratively ask

"why", exploring the cause and effect relationships (Ohno, 1988). Late 20th century root cause analysis (RCA) became formalized and widely adopted. It aims to discover the underlying causes of incidents (Rooney & Hauvel, 2004). With the emergence of systematic ways of looking at workplace safety, the emphasis has shifted from blaming individual workers towards addressing systemic causes of incidents.

The incident's course of events and the corrective actions taken are recorded in natural language. This poses a challenge for traditional statistical analysis methods that rely on structured numerical data. With the recent advancements in natural language processing (NLP), would it be possible to take a fresh approach to analyzing this underutilized textual data?

1.2 Applying NLP methods

In his mission to understand the genetics of guinea pigs, Sewal Wright formalized path analysis to explain the inheritance factor (Wright, 1921). The idea was to estimate the causal path coefficients by decomposing the correlations among a set of independent variables. As an end result the impact of certain genes causing certain phenotypes could be estimated. This analytical method is used as a foundation for developing causal inference methods, like Structural Causal Models (SCM) (Pearl, 1995) and PCMCI (Runge et al., 2019).

If one were to apply path analysis to textual safety data, the first step would be to identify the entities in the incident descriptions and the second step would be to infer a causal graph consisting of the entities and their relations.

Topic modeling is a common approach to grouping and labeling unstructured data. It has been used, for instance, in COVID-19 research to identify dominant topics in public discussion and to support the identification of possible root causes (Kokatnoor & Dr. K, 2022). In this research, it is used as a method for identifying themes in workplace safety incident reports. These identified themes can then be

treated as candidate nodes in causal chains and examined further using causal inference methods, in a manner related to Wright's path analysis. The purpose is to investigate whether such themes can help reveal plausible underlying causes of incidents. In probabilistic terms, the question can be expressed as $P(\text{Accident}_t \mid \text{Cause}_{i,t-k})$: what is the probability of an accident at time t given a possible earlier cause i at time $t-k$?

Or, when considering all possible candidate causes:

$$\{P(\text{Accident}_t \mid \text{Cause}_{i,t-k})\}_i \tag{1.1}$$

This can be read as the set of probabilities of an accident at time t conditional on different possible earlier causes. In order to make such an analysis possible, accidents and causes must first be represented as computable entities, after which causal relations between them can be evaluated.

1.3 Research questions and contributions

The object of the study is to move towards causal discovery in workplace safety data through an exploratory methodological approach. The hypothesis is that topic modeling can be used to identify meaningful themes in workplace safety reports and that these themes can serve as useful inputs for causal discovery methods. The thesis addresses the following research questions:

I Can coherent topics be identified from workplace safety data using topic modeling?

II Can the extracted topics be used to construct plausible causal relations that aid in interpreting and assessing the underlying causes and their impact?

The thesis proposes exploratory approach to causal discovery, utilizing topic modeling, large language models and causal discovery methods. The aim is to

examine whether topic modeling can be used to form meaningful themes to be used in modeling the workplace accidents with causal graphs, using Finnish occupational safety corpora. The thesis does not focus on performing the analysis using large language models, but utilizes LLMs in supportive roles.

1.4 Organization of this thesis

This introductory section presents the background and motivation for this research and theoretical basis for the methods used. The corpora and methods section explores the data used and the research methods are presented. The results and discussion section contains the results of the research and what can be concluded from that. The limitations of the research and further research opportunities are considered. Finally, the research is summarized with conclusions.

2 Theoretical framework

2.1 Occupational safety system data

Safety systems have practices in place to report and collect possible safety hazards, close calls, and accidents. This incident report is often followed by further safety analysis that determines the cause and addresses the issue either with suggestions or through direct corrective action. This process is reported and collected in the safety management system.

Root Cause Analysis (RCA) is a structured method used to examine incidents beyond their surface manifestations and to identify contributing factors and underlying causes. Although RCA procedures vary across domains, the literature commonly describes the process as involving collecting data, charting causal relations, identifying root causes, and proposing corrective actions (Rooney & Hauvel, 2004). Identifying the underlying root causes has the potential to advise operative and strategic decision-making related to workplace safety, which help reduce the likelihood of similar accidents recurring.

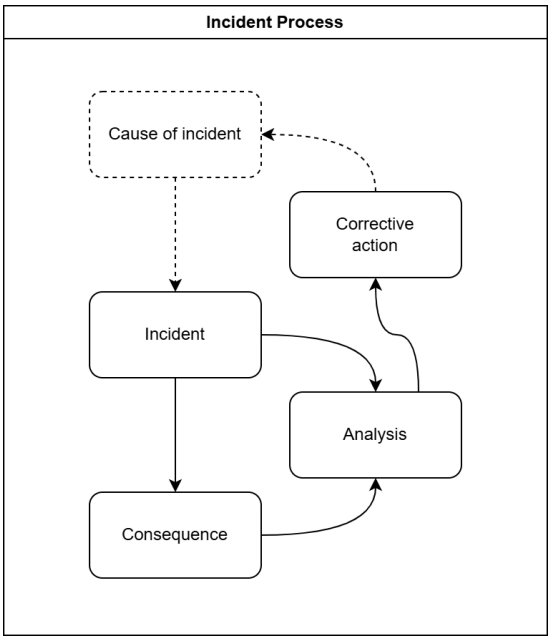

Figure 2.1: Typical flowchart of incident process. The cause of the incident is not directly observed but inferred, and the corrective action is recorded as a proposed response to the assumed cause.

2.2 Topic modeling

BERTopic is a compilation of different machine learning techniques coming together into a single pipeline for topic modeling. (Grootendorst, 2022) It has a modular structure which makes it possible to exchange its modules with alternatives. BERTopic consists of several modular steps that are typically applied in sequence. First, the input documents are converted into dense vector embeddings. These embeddings are high-dimensional semantic representations, with an exact count of dimensions depending on the embedding model used. These vector embeddings are then reduced to a lower-dimensional space to make the following clustering phase more manageable, with typically 2–10 dimensional components. Next, the clusters are named with a combination of bag of words, c-TF-IDF, and LLMs, to make them into interpretable topics.

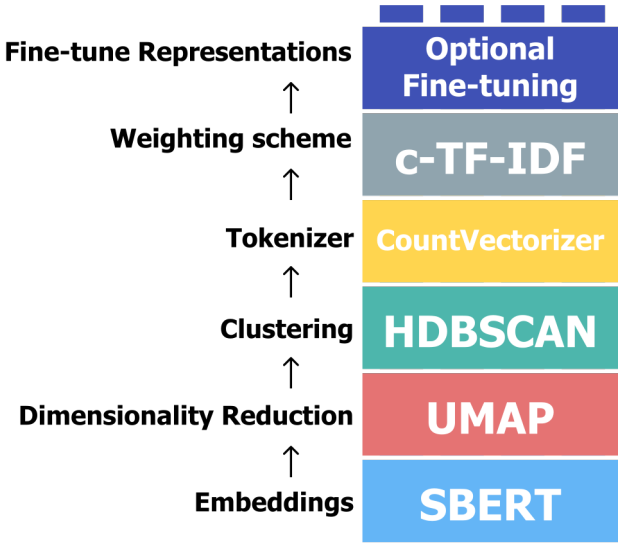

Figure 2.2: BERTopic algorithm with its default modular steps. Reproduced from the BERTopic documentation (Grootendorst, 2024).

2.2.1 Embedding with Sentence Transformers

For the purposes of topic modeling, each document needs a single vector representation in the common semantic space for the documents. One possible approach is to use a model like BERT or FinBERT to generate token embeddings and then derive a single document vector out of those embeddings, for example by averaging all the embeddings together. However, an approach like this is not optimal. A better alternative would be to produce an embedding directly from the document. Compared to word embeddings where each token in an input sentence is converted into a list of vector representations, a Sentence Transformer model encodes the whole input document into a single embedding, which represents the semantic meaning of the whole document. For the purposes of topic modeling, this is more suitable, since the aim is to cluster the documents by semantic similarity. BERTopic's default embeddings are generated using Sentence Transformers.

BERTopic is fully modular and by default it uses the Sentence Transformers pretrained model `all-MiniLM-L6-v2` for embedding the documents and `paraphrase-`

`multilingual-MiniLM-L12-v2` for multilingual. Other embedding models used for this thesis are `TurkuNLP/sbert-cased-finnish-paraphrase`, which is a 125M parameter model trained from FinBERT. It is specifically trained with Finnish paraphrase pairs (over 100 000) (Kanerva et al., 2023). `intfloat/multilingual-e5-large-instruct`, which is a 560M parameter model (L. Wang et al., 2024). The model is a top performer with under 1B parameters in the HuggingFace-maintained benchmark MTEB, for performance measurement of text embedding models. One of the tasks measured is clustering performance (Muennighoff et al., 2023).

Embedding models can be adapted with domain-specific data to improve performance on downstream tasks. A common unsupervised approach is continued pretraining with masked language modeling (MLM) on domain-relevant corpora in order to increase the model's familiarity with domain-specific language use. (Gururangan et al., 2020)

2.2.2 Dimensionality reduction with UMAP

Uniform Manifold Approximation and Projection (UMAP) is a general-purpose dimensionality reduction technique (McInnes et al., 2020). It is designed primarily to preserve local structure while also retaining some aspects of the global structure of the data, which is useful for clustering tasks with semantic embeddings. It was proposed in 2018 by McInnes et al. It utilizes manifold learning and is constructed from a theoretical framework based on Riemannian geometry and algebraic topology.

The UMAP algorithm constructs a weighted graph (which can be represented as a fuzzy simplicial set in topological terms) of local relationships by finding the `k` nearest neighbors for each node. The graph is then transformed into a lower-dimensional embedding by first initializing the data points as random positions in the target space and then iteratively optimizing their positions to minimize the cross-entropy between the original and the embedded lower-dimensional fuzzy simplical

sets. The goal is to encode the local structure in a lower-dimensional representation.

Due to the curse of dimensionality, the range of possible configurations is great, which causes clustering algorithms to perform poorly with high-dimensional data, like sentence embeddings. The distance metrics used in clustering become less meaningful. To correct that, dimensional reduction is performed with UMAP. The key hyperparameters that control the dimensionality reduction include:

- `n_components`: determines how many dimensions the target embedding has
- `n_neighbors`: determines how many neighbors are considered when learning the local structure
- `min_dist`: determines the minimum distance how tightly points are allowed to be packed together in the target embedding.

Unless a random seed is fixed, the UMAP algorithm is non-deterministic for two reasons. First, the initial positions in the low-dimensional embedding space are randomly set and, secondly, UMAP uses stochastic gradient descent when minimizing the cross-entropy objective. This causes different runs with the same hyperparameter configuration to give different results. If two different runs provide very different clustering results, this would indicate that the clustering is arbitrary and dependent on randomness. Because of this, the aim is to identify a configuration that produces consistent clusters. This stability can be estimated by comparing the resulting clusters with the adjusted Rand index (Steinley, 2004).

2.2.3 Clustering with HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is an unsupervised clustering algorithm. It is developed from DBSCAN, which forms groups from unstructured data (Campello et al., 2013). DBSCAN

struggles in identifying clusters with varying densities. This problem was addressed with HDBSCAN, which does not have this assumption of uniform density at the global level.

To briefly describe the main steps in HDBSCAN and how its key hyperparameters relate, first the core distances are calculated for each data object; this is the object's distance to its `min_samples`-th nearest neighbor. Then mutual reachability is calculated for each pair of data objects using both the core distance calculations and Euclidean distance (by default), resulting in a weighted graph. By computing a minimum spanning tree from this graph, HDBSCAN can construct hierarchical clusters from the tree. A cluster is only considered valid if it has at least `min_cluster_size` objects. Smaller clusters are pruned and their data objects will be assigned to other clusters or they are labeled as outliers.

2.2.4 Representation layers

Resulting clusters are interpreted independently from the previous methods that contributed to generating the clusters. Since HDBSCAN's clusters can have varying densities and shapes, these requirements need to be passed on to the representation method as well. In its default configuration, BERTopic generates a bag-of-words representation with sklearn's CountVectorizer, basically counting all the words in a cluster. All the clusters' bag-of-words representations are processed with c-TF-IDF to extract keywords from each cluster, c here representing class-based.

The resulting keyword list can then be sent to an LLM with a list of representative example documents for human-readable interpretation of the topics.

2.2.5 Topic cohesity with NPMI

To find the optimal values for UMAP and HDBSCAN, an evaluation metric is required. Normalized pointwise mutual information (NPMI) is a commonly used

metric for evaluating topic coherence in topic modeling (Lau et al., 2014; Röder et al., 2015). It quantifies how strongly pairs of words are associated based on their co-occurrence probabilities.

Pointwise Mutual Information (PMI) between words w_1 and w_2 is defined as:

$$\mathrm{PMI}(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)\,P(w_2)} \tag{2.1}$$

The result is normalized to range from -1 to 1. In topic modeling, higher NPMI values indicate that the top words of a topic tend to co-occur more often, which is generally interpreted as higher topic coherence. Lower or negative values indicate weaker semantic coherence among the topic words.

2.3 Causal discovery

The development of workplace safety practices has gone hand in hand with the development of statistical methods. Modern statistics began to take shape around the turn of the twentieth century. A major milestone in this development was Karl Pearson's formalization of the correlation coefficient in the late nineteenth century, which laid the foundation for one of the most widely used statistical measures for association (Pearson, 1896). Pearson's work contributed to a view in which statistical association became the primary rigorous way of describing relationships between variables, while causal interpretation was treated with greater caution. Basically, in this framework, causality is seen as a scientifically non-rigorous concept because it cannot be measured in the same sense as correlation can be measured. Judea Pearl argued that the strong emphasis on statistical association in twentieth-century statistics slowed the development of formal tools for causal inference (Pearl, 2009). As an alternative, he proposed a causality framework to study causality.

A central component of Pearl's framework is the directed acyclic graph (DAG), in which nodes represent variables and directed edges represent assumed causal effects

between them. DAGs provide a formal way to encode causal assumptions and allows identification of causal relations from statistical association alone using observation data. In this sense, the data can be considered to have latent causalities, and the task of causal discovery is to infer this structure from the available observations.

2.3.1 Granger causality

Granger causality is a statistical test proposed by Clive Granger in 1969 (Granger, 1969). The test evaluates whether one time series contains useful information to predict another time series.

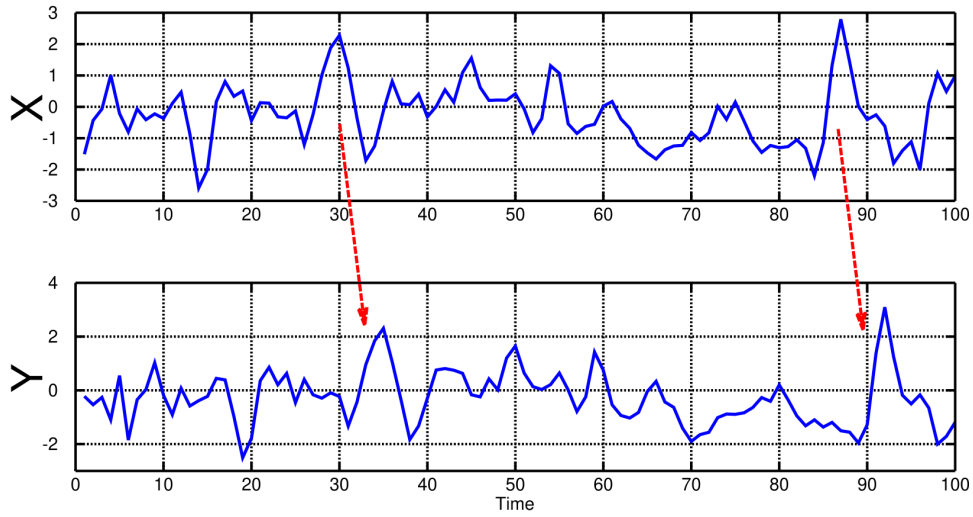

Figure 2.3: Illustration of when time series X events are repeated by time series Y after lag (BiObserver CC BY-SA).

The assumptions for the test are the following: 1. Future events are caused by prior events. 2. A cause does not have redundant predictive information, information explained by other variables.

The test compares a restricted autoregressive model of Y_t with an unrestricted model that also includes lagged values of X_t:

$$\begin{aligned} \text{Restricted form:} \quad & Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \epsilon_t \\ \text{Unrestricted form:} \quad & Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{j=1}^{p} \gamma_j X_{t-j} + u_t \end{aligned} \tag{2.2}$$

Y_t is the current value in the time series, which is the sum of three terms: the average value of the time series, the sum of lagged values, and an error term. If the lagged values of X significantly improve the prediction of Y compared to the restricted model, X is said to Granger-cause Y. This indicates predictive usefulness rather than causality in a philosophical sense.

2.3.2 PCMCI algorithm

Peter and Clark momentary conditional independence (PCMCI) is a robust causal discovery method for multivariate time series data. The algorithm estimates a causal graph from complex and noisy multivariate time-series data. PCMCI is based on (1) the PC-stable algorithm, where first all the potential parents are estimated for each variable by a series of conditional independence tests, and (2) the Momentary Conditional Independence test, where each potential link is tested while both the source and target variables are conditioned on the parents (Runge, 2024). PCMCI is implemented in Python by the Tigramite package that is used in this research.

PCMCI assumes that there are no unobserved confounders. To account for the latent variables, the options are to broaden the scope by introducing external variables like temperature, weather conditions, time of day, etc. Another option is to use a variation of PCMCI that introduces latent variables, LPCMCI. This allows for discovery of unobserved confounders affecting two variables.

3 Materials and methods

In this study, a method is presented that utilizes topic modeling to quantify incidents and preventive measures. The topic models are optimized for the used safety corpora. Each topic represents the incident or preventive measure cluster to which the description belongs. Mapping the frequency of topics occurring through time and place creates time series for each cluster. This opens up the possibility to determine if there is predictive temporal relationship, that is, a re-occurring temporal correlation where one time series follows another.

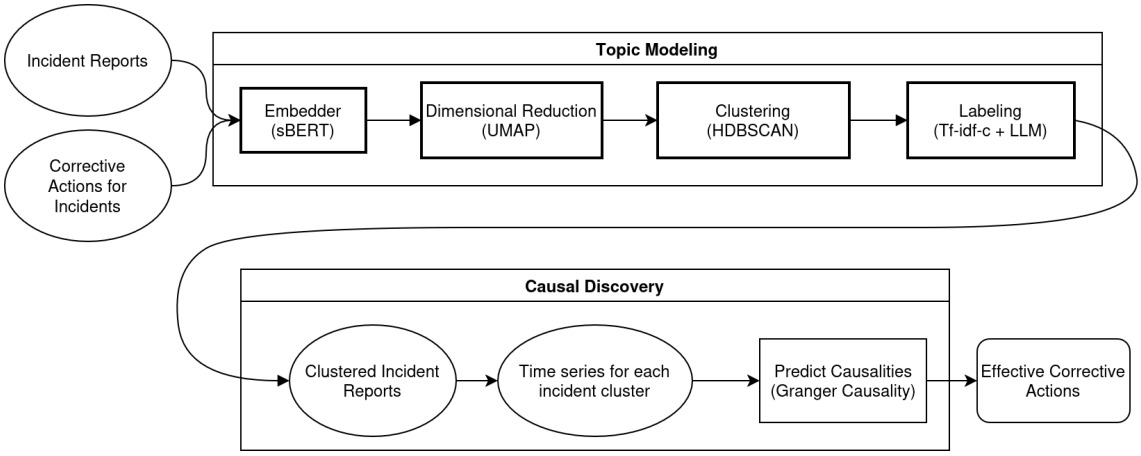

Figure 3.1: Overall structure of the proposed architecture.

In the topic modeling phase, the incident reports are embedded with a sentence embedder. These resulting vector embeddings are then reduced to lower dimensions with UMAP, after which they are clustered with HDBSCAN into topics. Finally, the

resulting topics are labeled with c-TF-IDF and a LLM. The details of these methods, how they are optimized for the data, and how their performances are evaluated are explained in the following sections.

3.1 Corpus

The primary dataset for this study is drawn from occupational safety reports collected in a context shaped by the European Parliament and Council Regulation (EC) No 1338/2008, which establishes a legal framework for the collection of statistics on health and safety at work (European Parliament and Council of the European Union, 2008). This corpus of workplace reports can be used to identify recurring patterns and support systematic analysis.

3.1.1 The origin of work safety data

The data used in this research originate from TAVA, an occupational safety management system developed by the software company Sofor. The system has been used by various companies across different domains, ranging from manufacturing to commerce. SSAB has granted permission for its historical data stored in the TAVA system to be used for this study and its publication. The data were provided for research purposes and are handled in anonymized form.

In practice, incident reports are often submitted by supervisory personnel, such as shift managers. The data is collected in order to improve safety measures through analysis and follow-up measures.

3.1.2 Exploration of data

The safety corpus spans a total period of 15 years from various locations and it contains several distinct report types. Five tables containing longer natural-language

descriptions were chosen from the TAVA corpora for the exploratory analysis phase: two types of incident reports, one for hazards/near miss situations and one for accidents; corrective action reports; deviation reports; and safety moments.

The dataset used for this research contains 6790 accident reports and 238 316 hazard reports and 193 941 corrective action reports. The incident reporting system is designed to capture information about the accidents, near-misses and safety measures that occur in the workplace.

Table 3.1: Correspondence between Hazard, and Accident reports

	Hazard report	**Accident report**
Cause of Incident	Why / How	Cause of Incident
Incident	What / Where	Description
Consequence	Consequence	Description
Analysis	-	-
Corrective Action	Immediate and Suggested Actions	Corrective action

Example of accident report (translated from Finnish):

> "The installers (working as a pair) were performing maintenance on an ERP (special heavy fuel oil) pump at a power plant. The installers loosened the pump casing, and shortly after loosening it, the casing shifted and hot ERP oil splashed onto one of the installers' face. There were small splashes across the face and a larger splash on the right cheekbone. In addition, there was ERP oil on the fingertips of the installer's right hand."

The temporal distribution is uneven throughout the reporting period. Accident reporting started in 1990, but the frequency before 2006 was so low that this period was excluded from the analysis. Hazard reporting started in 2006 and it dominates

the corpus throughout the observation window, while accidents occur at lower frequencies. The general trend is that hazard reporting has increased, possibly due to more comprehensive reporting practices, while accident reports have declined. This pattern may reflect improvements in safety, changes in reporting behavior, or both.

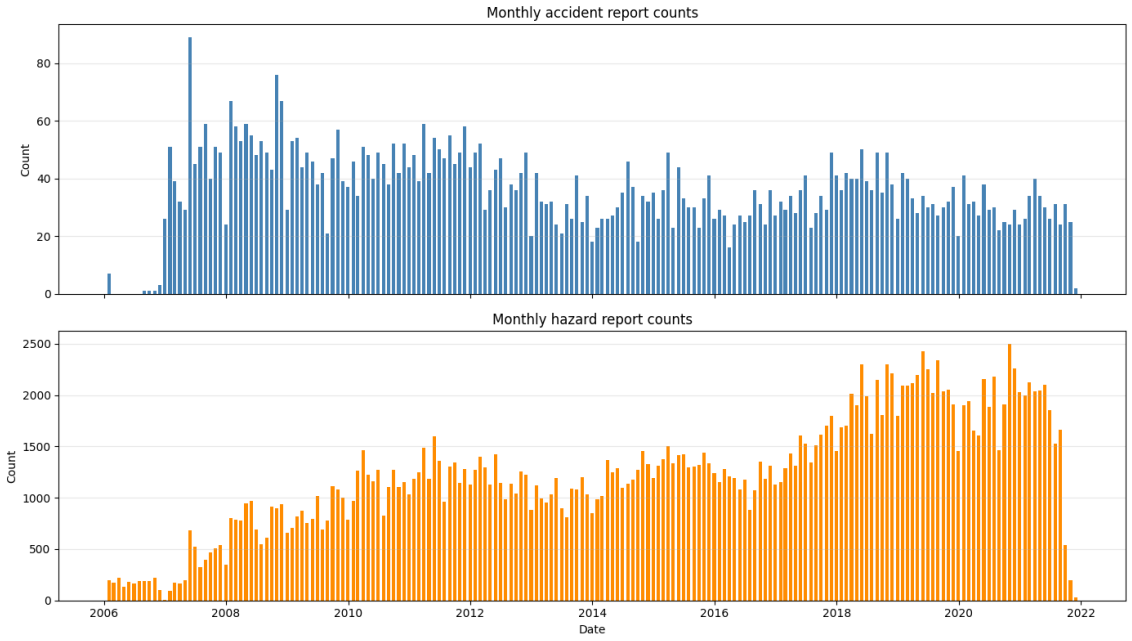

Figure 3.2: Monthly counts of accident and hazard reports over time from 2006 onward. The datasets are shown in separate panels due to their differing scales.

The corpus does not contain detailed investigation records of the incidents themselves. Additionally, in the accident reports, the consequence is not recorded as a separate field; instead, it is described as part of the incident field itself. Despite the comprehensive structure in the safety reports, only the incident description fields are mandatory, causing missing data in the other fields spanning from 35% to 83% of the reports, as described in the missingness table 3.3.

Example of hazard report (translated from Finnish):

> "The cold billet transfer car has a seat that can be used while driving the car. Chains have been installed in the access opening next to the seat as fall protection, but based on their condition they do not appear to be in active use (the chain had been threaded through the railing attachment loop, and trying to remove it from there was not quick). This creates a falling hazard if the seat tips over."

Table 3.2 summarizes the coverage of the natural language fields in the hazard reports. Only 5 % of the reports had all seven columns filled at the same time. If evaluation info is excluded and all values with 3 or fewer characters are dropped, the total is 30 475 (16,5 %). While there were multiple semantically relevant fields, their completion rates varied significantly. Only the incident description was filled consistently, while descriptions on the reported cause and consequences were often missing. This uneven coverage informed the decision to combine the textual fields together in the later processing phase.

Table 3.2: Coverage and missingness of the main textual fields in the hazard reports.

Textual field	**Count**	**Coverage (%)**	**Missing (%)**
Incident description	238 250	100.0	0.0
Cause description	59 412	24.9	75.1
Consequence description	64 911	27.2	72.8
Immediate actions	154 096	64.7	35.3
Suggested actions	71 875	30.2	69.8
Evaluation information	38 339	16.1	83.9

The accident reports contain 60 different columns, including accident description, information about personnel involved, investigation information and European Statistic on Accidents at Work (ESAW) and other classifications of the accident.

The table 3.3 summarizes the natural language fields in the accident data and their missingness ratio.

Table 3.3: Coverage and missingness of the main textual fields in the accident reports.

Field	**Count**	**Coverage (%)**	**Missingness (%)**
Locality information	6 179	91.4	8.6
Accident location description	443	6.6	93.4
Safety regulation deviation	689	10.2	89.8
Treatment location	4 005	59.3	40.7
Accident description	6 758	100.0	0.0
Work type description	6 441	95.3	4.7
Cause information	2 204	32.6	67.4
Immediate corrective action	2 525	37.4	62.6
Corrective action	4 520	66.9	33.1

The textual fields are on average short, with an average length of approximately 200 characters and most observations remaining below 750 characters long, after outlier removal. This is relevant for topic modeling, as short texts provide less context to utilize when separating topics from one another.

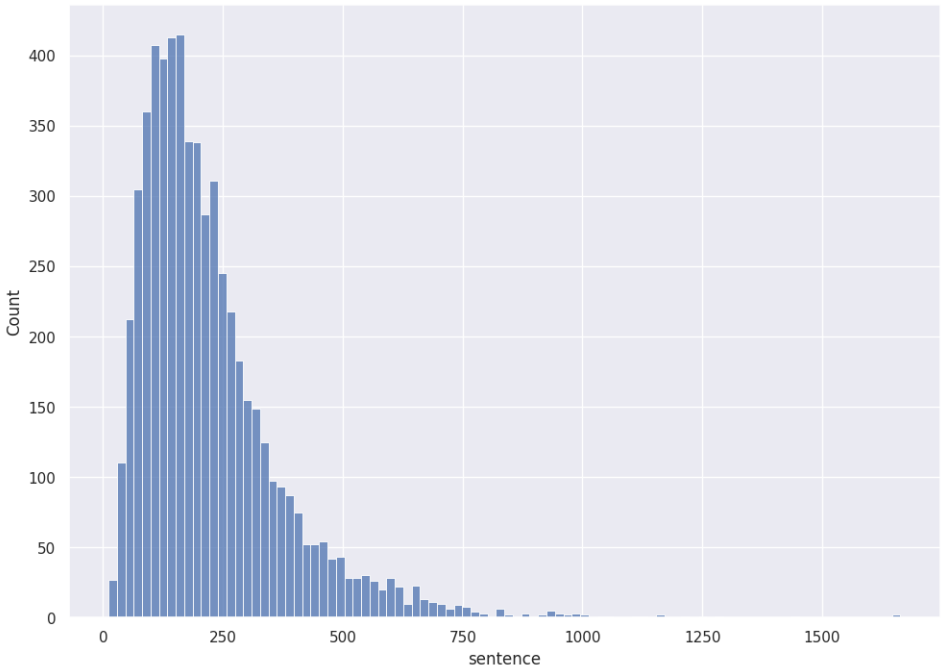

Figure 3.3: Sentence length distribution of accidents.

Based on the exploratory analysis, the selected text fields were those that contained high coverage of semantic relevance to the incident process described in Figure 2.1, namely the incident description, observed consequences, reported cause, and corrective actions taken. The safety moment and deviation report types provide additional context on observed deviations in the workplace and safety moments that are training moments for the employees. They are not utilized to keep the focus of this study on the frequency with which incidents occur.

Example of corrective action report (translated from Finnish):

> "More meters will be acquired, and the shifts will begin taking care of calibrating the meters themselves. Instructions already exist."

3.1.3 Preprocessing

For the topic and causal modeling, a subset of the fields was chosen for further processing. These included the date and time of the incident, the location of the incident, and textual descriptions of the incident. The textual descriptions covered what happened, the cause of the incident, and the actions taken to address it, such as first aid treatment, repair or replacement of equipment, and worker training.

Hazard and accident reports were filtered by removing rows whose main incident description fields were empty or did not contain alphabetic characters. Corrective action records were cleaned separately using the same criterion. Duplicate entries were also removed and all the reports written in all uppercase were lowercased.

The languages of the reports were detected with the `langdetect` library (Mimino666 & Nakatani, 2021). The majority of the reports were in Finnish (95%). Other languages like English, Polish and Swedish were dropped out of the dataset because they didn't represent a significant portion of the dataset.

The corrective actions are stored in their own table, and in the preprocessing phase, the corrective action descriptions are added to the incident rows with which they are associated. One incident report can have multiple corrective action reports related to it. These reports can contain mostly the same content, but with slight variations. All the corrective actions related to one incident are concatenated into a single string utilizing `difflib.SequenceMatcher` from the Python Standard Library (Ramos, 2026). SequenceMatcher detects the same and differing parts in two strings, which allows merging of two strings with variations together. When two conflicting variations were detected during merging, the longer string was chosen. Finally, all the orphaned corrective action reports that did not have any related incident were dropped from further analysis.

Location fields are free text with places written in a variety of forms with additional details about the location separated with "#". To normalize the written

forms in order to classify them, the locations are converted to lowercase, stripped and only the first part of the location, that is the municipality, is retained.

After standardizing the wording in the location fields, this revealed the uneven distribution across locations. In the accident data the two main locations accounted for 50.8% and 20.4% respectively, whereas the remaining locations only accounted for a small share. The hazard reports followed a similar distribution with the main location accounting for 62.8% and the second location 10.7% of the incidents. This imbalance indicates that reporting activity and recorded incidents were concentrated in a few locations, which informed the later focus on these places in the causal analysis.

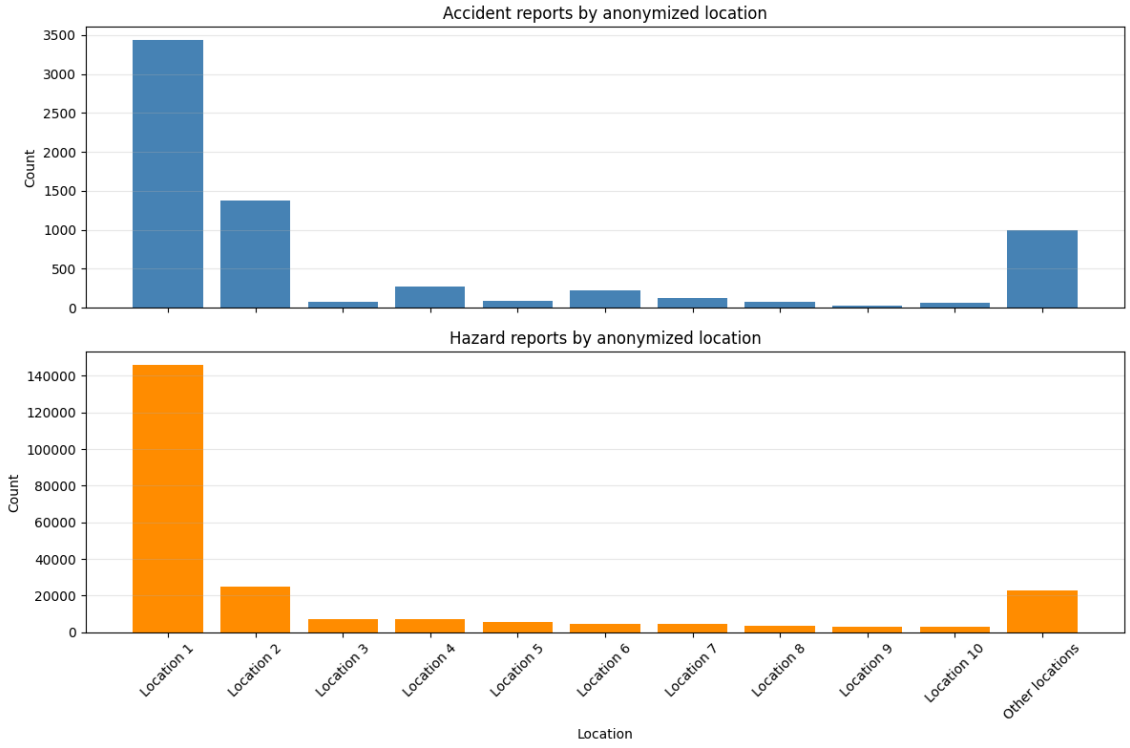

Figure 3.4: Distribution of accident and hazard reports across anonymized locations.

Since the hazard reports most frequently contained information in the "What/where" and "Immediate Action" text fields, this may indicate that the reports were often

filled soon after the incident and further evaluation was often left incomplete. Explicit causal reasoning from the dataset could be limited. Considering this, the cause, incident, and consequence columns are combined in the analysis to form a single textual representation for the topic model training.

3.2 Topic modeling

Topic modeling was performed for accident descriptions, hazard descriptions, and their respective corrective actions. For incident classification, BERTopic was chosen because it is an unsupervised method (Grootendorst, 2022), has been used in a variety of prior studies as a starting point for grouping unstructured text data (Madrid-Garcia et al., 2024; Nedungadi et al., 2025; Raman et al., 2024), and offers the potential for full automation of the process. This made it a suitable starting point for identifying semantically coherent clusters in the data without requiring predefined class labels.

Additionally, when the research was started in 2021, large language models had not been established as a robust method for zero-shot text classification as proposed by Z. Wang et al. (2023). For this reason, BERTopic was considered the more transparent and methodologically mature choice for this exploratory analysis.

3.2.1 Embedder selection

For the embedding model, three options were considered. `paraphrase-multilingual--MiniLM-L12-v2` is the default multilingual embedder in BERTopic, `TurkuNLP/sbert--cased-finnish-paraphrase` is a Finnish-specific sentence transformer, and `intfloat-/multilingual-e5-large-instruct` was the top multilingual performer on the MTEB leaderboard at the time of testing in April 2025 ("MTEB Leaderboard", 2025; L. Wang et al., 2024). All three models were evaluated by running the whole

pipeline with hyperparameter optimization.

All three embedding models were compared by running the full BERTopic pipeline, including hyperparameter optimization, separately for each model. The resulting topic labels were first assessed manually on a sample of 50 accident reports. After this preliminary comparison, the labels produced for the full accident corpus were evaluated with an LLM-based judge, Gemma 3 27b (Elangovan et al., 2025; Gemini Team et al., 2023). The purpose of this evaluation was to compare how interpretable the resulting topic labels were across the embedding models. The following prompt was used to determine which incidents had descriptive labels:

LLM Prompt to evaluate embedders

```
Given the accident description: {incident description}
Which of these labels are correct?
1. e5: {e5 label}
2. turkunlp: {turkunlp label}
3. minilm: {minilm label}

Return only:
1: Correct/False
2: Correct/False
3: Correct/False"
```

3.2.2 Clustering

Produced sentence embeddings are reduced with UMAP to a lower-dimensional space, and HDBSCAN finds the boundaries for the semantic clusters (Campello et al., 2013; McInnes et al., 2020). UMAP and HDBSCAN have several hyperparameters that affect cluster formation. To optimize the best set of hyperparameters, a

grid search strategy was used, with the end results being evaluated after running them through both UMAP and HDBSCAN. To reduce the number of parameter permutations, both processes were optimized independently with fixed parameters for the other process. A grid search strategy was used with the following values considered: optimized UMAP hyperparameters with values considered were `n_components` values 2, 3, 5; `n_neighbors` values 15, 30, 60, 100; and `min_dist` 0. For optimizing HDBSCAN, the optimal UMAP values were used, and grid search was performed with the following set of parameters: `min_cluster_size` values 5, 10, 15, 20, 30 and `min_samples` values 1, 5, 10.

For clustering sentence embeddings in higher dimensions, cosine similarity is the preferred choice for a distance metric (Reimers & Gurevych, 2019) because the direction captures the semantic similarity of the embeddings regardless of their magnitude. However, HDBSCAN uses DistanceMetric from scikit-learn, and support for that specific metric is not implemented. Instead, the Euclidean metric is used with normalized embeddings. In effect, this accomplishes a similar result, since for unit-length vectors Euclidean distance is closely related to cosine similarity. In effect, this makes Euclidean distance on normalized embeddings a practical approximation of cosine similarity.

For reference, for two unit vectors $\mathbf{u}$ and $\mathbf{v}$:

$$\|\mathbf{u} - \mathbf{v}\|^2 = 2\,(1 - \langle \mathbf{u}, \mathbf{v} \rangle) \tag{3.1}$$

where $\langle \mathbf{u}, \mathbf{v} \rangle$ is the cosine similarity between the vectors.

For evaluating the grid search, the cluster sizes were kept at a sensible size, meaning that configurations that produced only a few clusters or hundreds of clusters were discarded, and the best configuration was chosen as the one that had the best elbow detection score. For controlling the stochasticity of UMAP, the adjusted Rand index was employed for evaluating cluster stability (Steinley, 2004). Too unstable model configurations were ruled out.

3.2.3 Topic representation

To create a representation layer for the created topics, the documents in a cluster are converted into a single document, which is then lemmatized with Stanza to account for the many inflection points in the Finnish language. That document is converted into a bag-of-words representation and the words are ranked with c-TF-IDF (Grootendorst, 2022). The resulting top 5 words are used as keywords, and the documents are ranked based on how representative they are of the cluster (based on their proximity to the cluster in the embedding space). The top 5 representative documents, truncated to 200 characters, and the keywords are sent to the LLM model as a cluster naming task, with the following prompt:

LLM Prompt (original in Finnish)

```
Form a common topic from these keywords: {keywords}

Here's five samples where the keywords appear in context:
{documents}

Create a short, distinctive topic that represents the
   given keywords and documents, and is at most 4 words
   long.
The topic should describe the activity that caused the
   injury, taking the keywords into account.
For example: Do not name the topic slipping, but
   describe the circumstances that caused the slipping.

Return the answer in the format:
topic: {topic}
```

The LLM was instructed to label the topics using primarily the keywords generated by the bag-of-words and c-TF-IDF process. Five representative documents identified by c-TF-IDF are used as context.

For the LLM model, Gemma 3 27B was chosen because of its state-of-the-art multilingual support at the time of the research (Team et al., 2025), based on preliminary qualitative testing, it showed the strongest Finnish performance among the open-source models evaluated in this study. Other instruct models like Gemma 2 27B, Llama 3.1 8B, and Gwen 2.5 32B were tried as well, but their accuracy with Finnish grammar was significantly worse than that of Gemma 3. Because of the sensitive nature of the data, using external LLM services was not included in the comparison.

3.2.4 Semantic structuring

A key methodological challenge in unsupervised clustering of safety reports is that the clustering can happen based on any semantic features, such as descriptions of injury or other consequences of the incident. This can be problematic when the task is to support causal discovery, since clusters formed around similar outcomes can have very different causal mechanisms (Leveson, 2012). As a result of this directionless process, someone being injured is as equally important a feature as the description of how they were injured. In the thesis, a solution that was attempted to solve this was semantic structuring with an LLM.

Several alternative strategies were considered for reducing the influence of reports where consequences played a big part. One option would have been to segment the report into sentences and exclude the sentences that described the outcome. A more direct, surgical approach would have been to modify the embedding representation itself by first identifying the semantic dimensions related to consequences and then suppressing them. Direction manipulation of the embedding space was considered

too complex a task, since the unwanted semantic features do not align cleanly in their own dimensions.

Instead, sentence segmenting seemed promising, but since the identification of the sentences describing consequences required semantic evaluation, the approach would require the use of large language models for restructuring and could be taken further. The accident report could be converted into context, incident, and consequence parts, but instead a more straightforward approach was chosen that transforms the text to key identifiers of the incident. This is because running LLMs is computationally expensive and the incident reports tend to contain irrelevant information.

As established, the reports are filled in a variety of ways and they often contain information that goes strictly beyond describing the accident. This will naturally affect the resulting sentence embeddings. Intuitively, it would make sense that if one were to have cleaner texts, free from irrelevant text, and leave only the relevant content, the resulting embeddings would be better as well.

To preliminarily examine this idea, core features of the reports were extracted using Gemma 3 27b to get features like object, condition, and activity. An ad hoc structured extraction schema was developed for this study to represent core incident features in a compact format.

As an example, the following safety report is transformed to the following form: "Poltettiin työntöpuolen nousuputkea ovenkautta. Kuumia kipinöitä lensi suojahanskojen suuaukosta takin hihan alle. Useita pieni palovammoja vasemmassa kädessä."

JSON Result: `{'Object': 'Nousuputki', 'Condition': 'Palaminen', 'Activity': 'Hitsaus', 'Action_Taken': 'Ei toimenpiteitä kuvattu', 'Resolution_Status': 'Unresolved', 'Severity_Level': 3}`

This keyword extract is then converted into sentence embeddings and processed with the same downstream pipeline as the original pipeline with free-text reports,

namely dimensional reduction with UMAP followed by clustering with HDBSCAN. Hyperparameter tuning was not performed for the semantic structuring model due to time constraints. Instead, a hyperparameter configuration was chosen that produced a similar amount of clusters. This made it possible to get a preliminary comparison of whether the chosen semantically structured representations produce more causally relevant topic clusters than clustering based directly on the original report descriptions.

LLM Prompt

```
You are an expert Safety Data Analyst specializing in
   Finnish industrial environments. Your task is to
   extract structured causal factors from unstructured
   Finnish safety reports.

Output Goal:
Extract specific entities into a JSON format. You must
   convert all extracted words into their **Finnish Basic
   Form (Nominative Singular / Perusmuoto)**.

Schema:
1. "Object" (Kohde): Physical item/machine (e.g., "Trukki
   ").
2. "Condition" (Vaara/Olosuhde): The defect/hazard (e.g.,
   "Öljyvuoto").
3. "Activity" (Toiminta): Worker task (e.g., "Hitsaus").
   Return null if unknown.
4. "Action_Taken" (Toimenpide): Immediate action.
5. "Resolution_Status" (Tila): "Resolved" (fixed) or "
   Unresolved" (temporary fix/pending).
6. "Severity_Level" (Vakavuus): Integer 1-5.

Output only valid JSON.
```

3.2.5 Topic coherence

To measure the topic consistency score, all the documents were preprocessed by lowercasing the text and removing the most common Finnish stop words using the Finnish stopword list from the Spacy library. After this, the text was lemmatized. Several tools were considered, namely Trankit Finnish-TDT (Nguyen et al., 2021), Stanza (Qi et al., 2020), and Spacy (Honnibal et al., 2023). Trankit was abandoned after version dependency issues. Stanza was the more linguistically detailed model of the two remaining due to its neural pipeline, but it also introduced heavy resource usage. Considering the task of using lemmatization as part of the benchmarking process, faster speed was a more significant factor, which is why lemmatization was performed with Spacy.

For the topic coherence score, the NPMI method was used (Lau et al., 2014; Röder et al., 2015). NPMI was measured with unprocessed text, with lowercasing and removing stop words, and with full preprocessing including lemmatization. Lemmatizing the text ended up lowering the NPMI score, so it was not used. Lowercasing and removing the stop words increased the score slightly (0.055 -> 0.075), so that was chosen for preprocessing.

After preprocessing, each topic is iterated over and the ten most common words are found for each category. This word list is then used to measure NPMI consistency as described in section 2.2.5.

3.3 Causal discovery

3.3.1 Exploratory analysis with Granger causality

With the topic modeling complete, the topics can now be used for causal discovery. For the Granger causality analysis, the topic counts in the incident dataset were first aggregated into a weekly time series. For each week and topic label, the number of

observations was counted, forming a time series. Next, pairwise Granger causality tests were carried out between all accident topic time series and all corrective action topic time series, using lags up to five time steps. The idea is that if certain corrective action topics are followed by reductions in accident topics, this may indicate that the measures are associated with improved safety outcomes.

This approach should be regarded as an exploratory screening method rather than a confirmatory causal analysis. The topic count time series were sparse, highly discrete, and potentially affected by common trends, seasonality, and reporting intensity. Additionally, a large number of pairwise tests were performed, which increases the risk of false positive findings. For these reasons, statistically significant Granger links were interpreted cautiously.

3.3.2 Causal discovery with PCMCI

Hazard and accident topic data were aggregated into weekly time series. The data were grouped by site location, and for the analysis only the largest location was used (later referred to as Location A), encompassing 52.5% of the occupational safety reports. To reduce data sparsity, topics for which less than 15% of the weekly entries were non-zero were filtered out. The time series variables were standardized with z-scores. PCMCI was then run with a maximum lag of $\tau_{max} = 3$, allowing lagged dependencies at one, two, and three time steps. In addition, lag-1 autoregressive links were explicitly allowed for each variable. Partial correlation was used as the conditional independence test, with a PC selection threshold of $pc\_alpha = 0.2$, and a final significance level of $p < 0.05$ with Benjamini–Hochberg false discovery rate correction.

Table 3.4: Excerpt of the weekly topic-lag count matrix used as input for PCMCI.

Topic and Lag	**Week 1**	**Week 2**	**Week 3**	**Week 4**
Occupational injury outcomes Lag 1	6	1	5	6
Occupational injury outcomes Lag 2	1	1	1	5
Occupational injury outcomes Lag 3	3	1	2	1
Eye injuries from foreign objects Lag 1	0	1	0	4
Eye injuries from foreign objects Lag 2	2	0	3	0
Eye injuries from foreign objects Lag 3	0	1	0	3
Movement-related accidents Lag 1	0	1	1	1
Movement-related accidents Lag 2	1	1	0	1
Movement-related accidents Lag 3	2	2	2	0
Machinery-related hand injuries Lag 1	1	0	0	0
Machinery-related hand injuries Lag 2	0	1	1	0
Machinery-related hand injuries Lag 3	0	1	2	1
Commuting bicycle accidents Lag 1	0	0	2	3
Commuting bicycle accidents Lag 2	1	0	0	2
Commuting bicycle accidents Lag 3	1	0	0	0

Listing 1 Sparsity-Constrained PCMCI Causal Discovery

procedure PCMCI-DISCOVERY($D, \tau_{\max}, \alpha$)
    Extract hazard and accident variables from D
    $V \leftarrow V_{\text{HAZ}} \cup V_{\text{ACC}}$
    Remove rows with missing values
    Retain variables with non-zero proportion > 0.15:
        $V^* \leftarrow \{X \in V \mid \text{nonzero\_fraction}(X) > 0.15\}$
    if $|V^*| < 2$ **then**
        return empty graph
    end if
    Standardize each variable in V^* using z-score normalization
    $V^*_{\text{HAZ}} \leftarrow V^* \cap V_{\text{HAZ}}$
    Initialize allowed link set L
    for each $X_j \in V^*$ **do**
        Add autoregressive link $X_j(t-1) \to X_j(t)$ to L
        for each $X_i \in V^*_{\text{HAZ}}$ **do**
            for $\tau = 1$ to $\tau_{\max}$ **do**
                if $X_i \neq X_j$ **then**
                    Add lagged link $X_i(t-\tau) \to X_j(t)$ to L
                end if
            end for
        end for
    end for
    Run PCMCI with partial correlation using allowed links L
    Apply FDR correction to the resulting p-values
    Retain significant links:
        $E \leftarrow \{X_i(t-\tau) \to X_j(t) \mid p_{\text{corr}} < \alpha\}$
    return $G = (V^*, E)$
end procedure

4 Results

This chapter presents the results of the proposed solution architecture, which consists of several interdependent steps. Since the performance of the subsequent steps is dependent on the preceding stages, particular focus is on the early steps, i.e. topic modeling. The topics generated are in Finnish, but for brevity's sake they are translated into English and presented only in this translated form, unless the Finnish presentation is important to bring up.

4.1 Topic model selection

4.1.1 Topic model tuning and evaluation

Topic model selection consisted of choosing the embedder and hyperparameters for UMAP and HDBSCAN. The embedder was chosen based on manual examination and automatic detection.

In the manual examination, topics were evaluated qualitatively. For example, "The person was in a field group during a loading situation and when he came to take a break, he felt debris in his right eye." (translated from Finnish) was labeled as "eye injury" by the e5 embedder pipeline and "pipeline handling" by TurkuNLP. The first one would be considered correct and the second one incorrect. The sample could have multiple correct topics describing different aspects of the text; in this case there would be multiple different types of accurate answers.

The results of the test were the following: e5 had 22 correct, MiniLM had 20 correct, and TurkuNLP had 14 correct. TurkuNLP had the most outliers (23.8%), which contributed to the lower number of correct labels. The e5 model had only 18.2% outliers. This test was not rigorous, but it gave some preliminary understanding of the embedders. Reviewing the answers, it also became evident that often the clustering choices were on orthogonal axes. For example, one embedder had clustered based on the consequences and another by the type of job being performed. While the TurkuNLP embedder had fewer correct topics, it sometimes seemed to understand nuance that the other embedders missed.

After running through 50 examples, e5 seemed to provide the most accurate topics, reflecting some main aspect of the description, so it was chosen as the embedder.

A second, more thorough evaluation was performed with the LLM Gemma 3, instructing the model to classify the correct topics. All accident descriptions and their respective candidate topic labels were run through it. The e5 embedder got 60.5% correct, TurkuNLP got 57.7% correct, and MiniLM got 50.8% correct. The results supported the previous finding that E5 seemed to have the most accurate classification.

Reducing the embedding dimensionality to 2 and 3 components proved to be too stochastic as a result of adjusted Rand index stability tests. After ruling those options out, only 5 and 10 components were considered for the final results.

The hyperparameters for UMAP and HDBSCAN were chosen with grid search. The optimal values were chosen as a compromise between cluster size and topic entropy. The optimal values for UMAP were neighbors 15 and components 5, and for HDBSCAN min cluster size 30 and min samples 1.

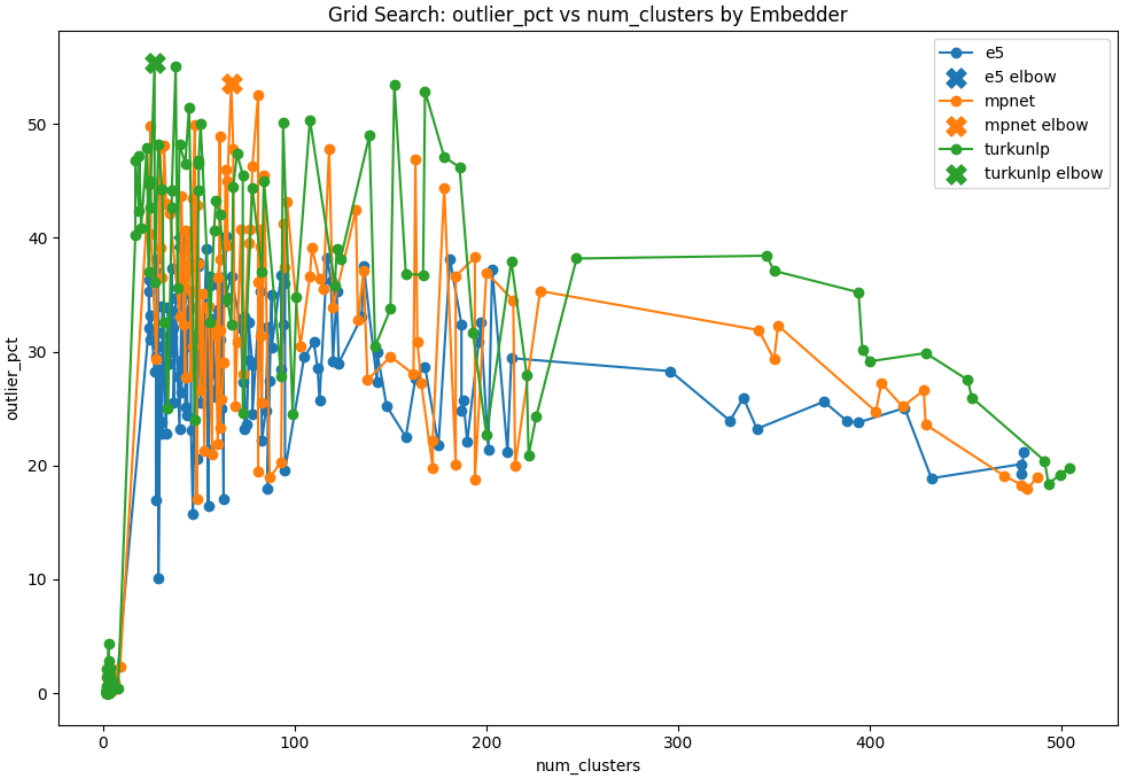

Figure 4.1: Results of elbow detection. Mpnet was chosen for this comparison instead of MiniLM. e5 provides the fewest outliers in all cluster counts.

The e5 embedder seemed to consistently provide the fewest outliers regardless of the hyperparameter setup.

After dropping configurations with the smallest clusters, cluster size and outlier share had a correlation of -0.32, suggesting that more refined clusters tended to have fewer outliers on average. Topic entropy and number of clusters were correlated at 0.91. Increasing UMAP dimensions seemed to increase outliers (0.40). With more dimensions, there is more freedom to position the points so that local neighborhoods are preserved.

Comparing the different embedders, TurkuNLP seemed to have the least topic entropy (average of 3.45).

To tune the hyperparameters, the accident corpus was used with its 6000 entries. Initial tests to tune the hyperparameters were run with grid search. The cluster sizes varied from 2 to 500. Outlier share varied from 0 to 58%.

Silhouette score was used as a metric to determine the optimal values for UMAP and HDBSCAN, respectively, for each sentence embedder.

Table 4.1: Optimal HDBSCAN and UMAP values for the e5 embedder.

Components	Neighbors	Min Cluster Size	Min Samples	Clusters	Outlier %	Silhouette
3	60	30	10	31	38.3%	0.696
3	60	20	10	40	38.3%	0.690
5	30	30	10	33	36.3%	0.688
5	15	15	10	53	34.7%	0.688
3	30	30	10	35	34.3%	0.687
5	60	30	10	28	38.6%	0.685
5	60	20	10	36	36.7%	0.682
5	15	30	10	38	32.4%	0.679
3	30	15	10	51	34.8%	0.677
5	60	15	10	43	36.1%	0.675

The NPMI coherence score for the chosen model with accident data was modest but positive at 0.075. An NPMI value above 0 indicates a positive statistical dependence between terms. The result exceeds the baseline for random noise in a topic modeling context as validated by Röder et al. (2015).

Table 4.2: Optimal HDBSCAN and UMAP values for the TurkuNLP embedder.

Components	Neighbors	Min Cluster Size	Min Samples	Clusters	Outlier %	Silhouette
5	30	10	10	42	49.5%	0.521
3	100	20	10	27	48.7%	0.499
2	15	5	5	216	37.1%	0.483
3	30	5	5	169	51.3%	0.474
2	30	5	5	225	40.8%	0.462
2	15	15	5	86	35.6%	0.461
2	15	10	5	123	33.0%	0.459
5	15	5	5	179	48.5%	0.456
2	30	20	5	62	35.8%	0.455
5	60	20	10	25	46.6%	0.447

Table 4.3: Optimal HDBSCAN and UMAP values for the MiniLM embedder.

Components	Neighbors	Min Cluster Size	Min Samples	Clusters	Outlier %	Silhouette
3	15	5	10	113	45.5%	0.577
5	15	15	10	61	42.2%	0.564
5	60	5	5	2	0.0%	0.564
5	15	10	10	74	42.5%	0.563
5	100	30	10	2	0.0%	0.560
5	30	15	10	55	47.4%	0.555
5	60	10	10	62	49.8%	0.550
5	60	30	10	2	0.0%	0.549
5	100	5	10	2	0.0%	0.549
5	60	15	10	45	51.0%	0.547

4.1.2 Topic modeling results

Topic modeling was performed for both accident descriptions and hazard descriptions, and respectively for their corrective actions, resulting in 4 topic models. The topic modeling pipeline started from sentence embedding and ended in labeling the topics.

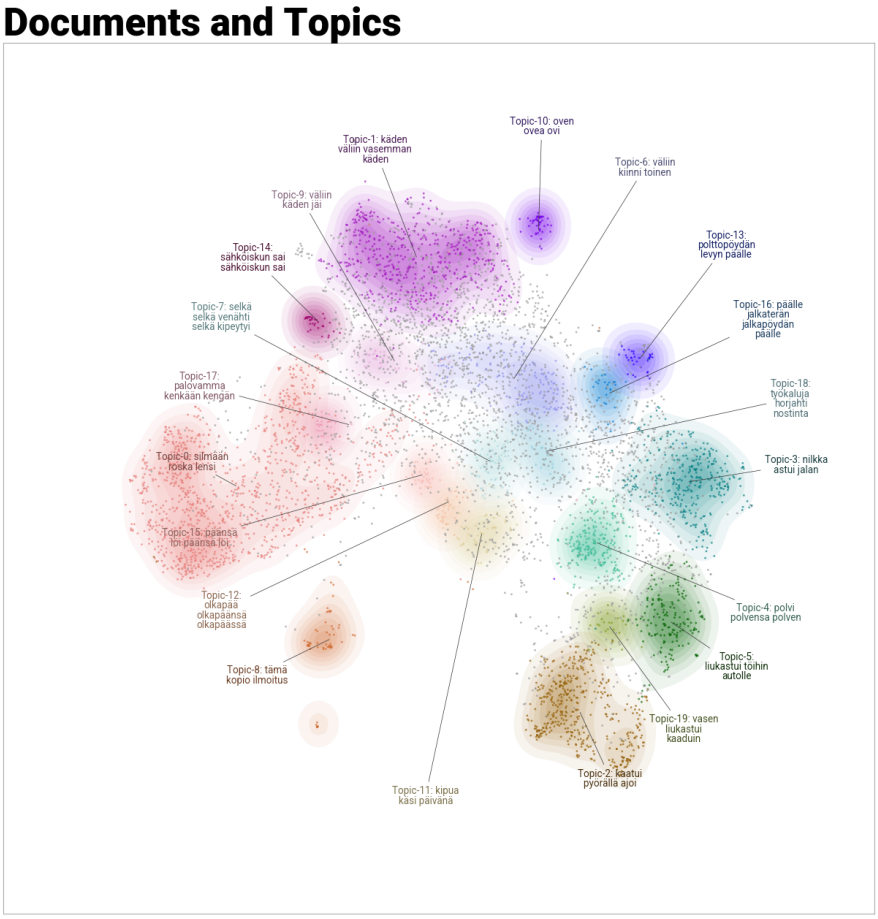

Figure 4.2: Accident clusters visualized in two dimensions using UMAP.

The task of topic modeling is to identify the type of accident that is happening. However, the accident description often also contains information about the consequences of the incident. For example, slipping on a staircase can cause a head injury or a leg injury, and it is possible that this kind of unintentional classification happens based on the wrong components of the data.

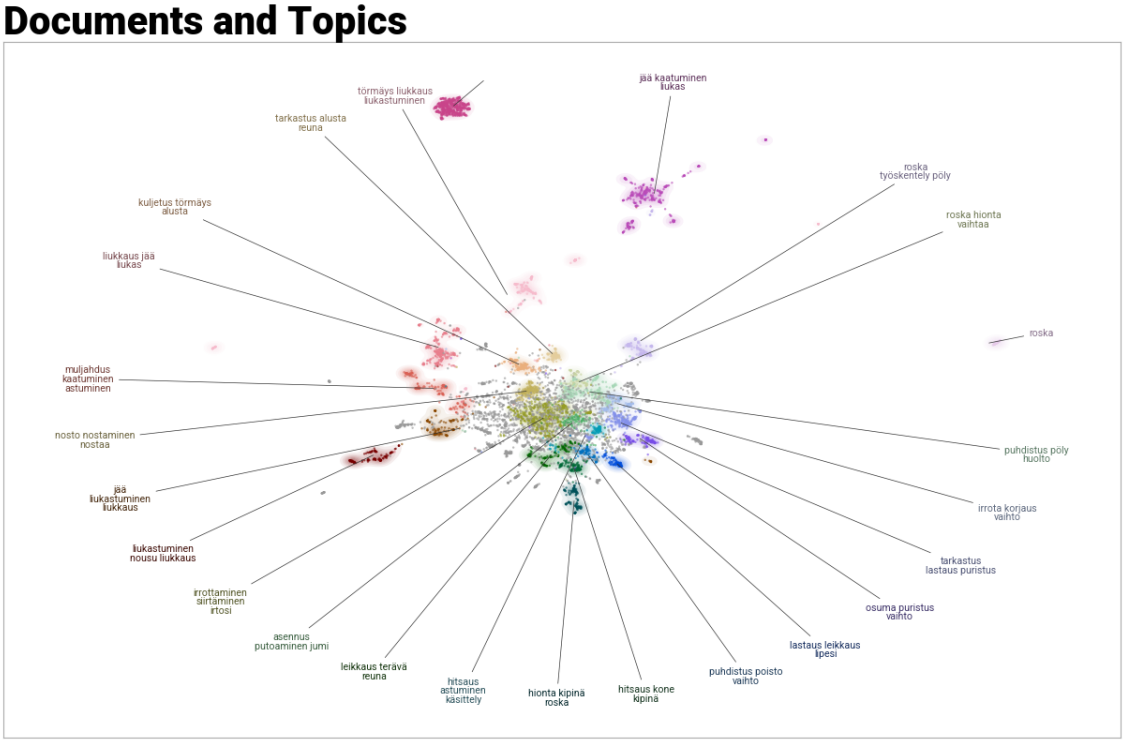

Figure 4.3: Accident clusters from the semantic structuring experiment, visualized in two dimensions using UMAP.

The semantic structuring experiment produced more visually distinct clusters with clearer separation in the two-dimensional UMAP projection. However, this did not translate into improved causal discovery results.

4.2 Causality results

4.2.1 Granger causality results

Granger causality tests were performed by comparing the clustered accidents with corrective actions. The tests were performed both at the company-global level and at the location-specific level.

The accident time series was sparse, and the smaller accident clusters contained

too few observations for statistically reliable testing. That is why the larger incident data and the larger mean values in the clusters were important for performing the causality test in a meaningful manner.

Table 4.4: Sample of the top 10 global results from applying Granger causality to workplace accidents and their corrective actions.

No	Accident	Action	p-value
1	Knee injury in an occupational accident	Electrical work safety and investigation	1.15×10^{-7}
2	Occupational accidents caused by flying objects	Occupational safety training and measures	2.18×10^{-7}
3	Falling and tripping on stairs	Handling of sharp objects	1.36×10^{-6}
4	Shoulder pain during lifting work	Protective equipment for welding work	3.48×10^{-6}
5	Rear-end collisions at intersections	Management of product defects and incidents	3.52×10^{-6}
6	Shoulder pain during lifting work	Management of product defects and incidents	6.18×10^{-6}
7	Tools falling onto feet	Paying attention to hazardous situations	2.00×10^{-5}
8	Handling of tools and parts	Finger crushing hazard during handling	3.75×10^{-5}
9	Hand crushing hazards at work	Installation safety of mechanical parts	3.86×10^{-5}
10	Handling and removal of scrap metal	Safety in cutting steel bands	5.30×10^{-5}

Many statistically significant associations were difficult to interpret causally, which supports the use of Granger causality here as an exploratory screening tool rather than as evidence of direct causal effects.

4.2.2 PCMCI results

PCMCI was applied to the weekly topic-frequency time series from the largest location (later referred to as Location A), encompassing 52.5% of the occupational safety reports. Using the baseline topic model, six significant lagged links were identified after Benjamini–Hochberg correction at the $p < 0.05$ level. All detected links were between topics with incident type hazard, while no significant links involving accident topics were found. Figure 4.4 visualizes the resulting network of significant lagged dependencies.

One of the six significant PCMCI links connected the hazard topic *Load on snow barriers and canopies* to *Condition and environment of the mixer* at a lag of three

weeks. Inspection of representative reports suggested that the source topic mainly captured winter-related incidents such as snow accumulation, slippery surfaces, and falling snow, whereas the target topic captured a variety of mixer-area-related safety issues, such as equipment condition or access to emergency supplies. The examples did not indicate any obvious mechanism linking the two topics together, but some non-obvious latent mechanism is possible.

Another link was found between *Blast furnace gas leak and carbon monoxide* and *Overflowing contents in waste container*. Blast furnace disturbances often trigger urgent repairs, temporary arrangements, and increased use of consumables. As a result, the activities produce an unusual amount of waste, which can cause overflowing bins and waste left in the work areas.

Perhaps the most probable link was between *Deficiencies in burner condition* and *Obstructions and blockages on table*. Poor condition of the cutting machine can cause interruptions, abnormal material flow, and manual intervention. Unfinished material and tools then accumulate on nearby tables, and workers are forced to improvise around the disruption.

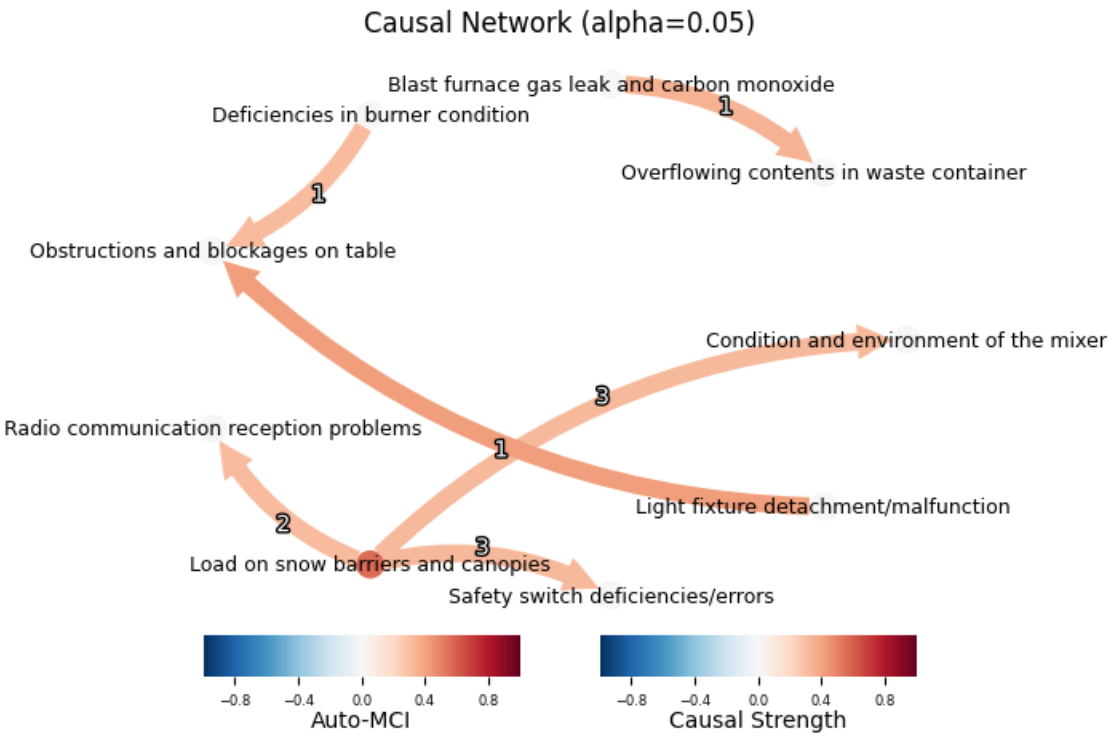

Figure 4.4: The found causal links from hazard to hazard visualized as node graph.

In addition to the baseline topic model, PCMCI was run using the topics generated with the LLM-assisted semantic structuring. The experiment did not yield any statistically significant links for Location A after Benjamini–Hochberg correction.

5 Discussion

This chapter is dedicated to reflection on the study. It consists of interpreting the results, considering what the results imply, discussing applications of this study, identifying challenges that were encountered during this study, and outlining possible future directions that could follow from this study.

5.1 Interpretation of the results

There were two main areas in the study: detecting incident entities with topic modeling and inferring causal relations between these entities based on the frequency and timing with which they occur in the corpora.

The topic modeling consists of a series of sequential methods running one after another. Some of these methods were evaluated individually on their own, but for many components it was difficult to pinpoint exact evaluation metrics, and instead they were evaluated as part of the whole pipeline using the end results. The embedders were evaluated with LLMs, with the results indicating that the choice of the embedder is significant in affecting topic modeling accuracy. A notable result was the evaluation method itself with the Gemini 3 LLM, which proved to be effective and in alignment with human judgment. Its use markedly reduced the required human labor in evaluating performance. This approach builds on the methodology of Lau et al. (2014), who demonstrated that automated metrics can effectively emulate human judgment in topic interpretability. The choice to use Gemini 3 for label

evaluation aligns with the critique by Hoyle et al. (2021), who argue that traditional coherence measures are often insufficient for modern topic models.

The top two performers in the LLM-based label evaluation were the E5 embedder (60.5%) and the TurkuNLP (57.7%) embedder. The first one is a 560M-parameter multilingual model and one of the top performers on the MTEB leaderboards at the time of testing, and the second one was a 125M-parameter model specifically fine-tuned with Finnish corpora for sentence embeddings. When working with Finnish-language data and considering the smaller size of TurkuNLP, it would be easier to fine-tune it with domain corpora. A 560M-parameter model would impose limitations on batch size, maximum token length, and precision determined by the available VRAM for training. Setting precision to fp16 does not impact accuracy (Micikevicius et al., 2018), and a smaller batch size makes the training slower, so only the maximum token length has the potential to impact accuracy.

During the study, the TurkuNLP sbert model encountered more instability. Runs with the same hyperparameter configuration, controlling the UMAP random state but with slightly different data, sometimes produced 2 clusters and other times 50 clusters (McInnes et al., 2020). This behavior was not witnessed with the other two embedders. This could be an indication of some underlying problem in the model or of an undetected factor in the testing setup.

There is a limit to how much the embedder choice itself can improve performance. The subsequent clustering and representation stages play an important role. When choosing the right hyperparameters, one has to make judgment calls about the correct cluster count, topic entropy, and outlier size – not to mention the qualitative impression the clusters leave when examining them. Reducing the outlier size would be an obvious goal, as a large portion of the data remains unclassified and thus excluded from the generated topics (Campello et al., 2013). In theory, cluster count should not be a problem that cannot be overcome with some effort, since clusters

can always be merged together, which makes it more of a heuristic problem. For the purposes of generating dense time series data, there cannot be too many clusters. Then again, too few clusters would affect the granularity of the data and very semantically differing incidents would end up being clumped together.

The evaluation method used for choosing the embedder could in theory be applied to the hyperparameter grid search, but the process would have to be made more practical through optimization. This could be done with methods like sampling, batching, and splitting the grid search and running it for each phase separately in the topic modeling pipeline.

Utilizing Gemma 3 for generating labels for the clusters performed above expectations. It was the most performant open-source model for Finnish among those considered in this study.

Although the Granger causality framework identified several statistically significant associations between accident and corrective action topic clusters, closer semantic inspection showed that many of these pairs were open to interpretation and did not support firm conclusions. This suggests that, in this setting, pairwise Granger testing was more effective as an exploratory detector of temporal covariation than as a reliable identifier of actionable causal structure. This limitation is well-documented by Eichler (2008), who notes that latent variables in multivariate time series often lead to spurious causalities.

Running PCMCI discovery with the base model and discovering 6 significant links showed that this approach could be viable, leaving room for further optimization. As it is, the sparsity of the constructed network suggests that only limited temporal structure indicating causal links could be formed from the weekly topic series. The topic series contained long sequences of zero observations, especially for smaller topics, while the most frequent topics appeared at modest weekly counts. As noted by Runge et al. (2019), the statistical power of a method like PCMCI to

detect robust lagged dependencies is limited in such conditions.

The absence of significant links in the semantic structuring experiment suggests that restructuring the topic representation did not improve the recoverability of temporal causal links in this subset of the data. It is possible that the chosen approach was too heavy-handed, removing too much information for the sentence transformer. The alternative research idea of segmenting the text into three parts – context, incident, and consequences – could be worth revisiting. So while the preprocessing with an LLM did not provide any value with the chosen prompt, focused prompt optimization would be necessary before drawing conclusions about whether LLM preprocessing could provide additional value for the process.

5.2 Practical implications

The methodology developed in this research was applied to the analysis of occupational safety data from two separate companies during autumn 2025 and winter 2026, respectively. The findings of the analysis were presented to the project's steering group. The resulting analysis has been utilized in forming a situational picture of the possible causes of serious safety issues. The analysis also dispelled worries about unnecessary incident reporting, underlining the importance of good data collection practices and stressing how comprehensive data collection, as presented in Figure 2.1, is a fundamental first step for further analysis of the root causes of problems.

An implementation of the methodology is also planned for TAVA's occupational safety system. The purpose is to provide real-time analysis for safety administrators as data is being collected and to provide the possibility for early intervention. Further development of the causal process is necessary to recognize early signals. This approach aligns with the principles of proactive safety culture and systems thinking, as advocated by Leveson (2012). One possible method is to generate synthetic data

to boost the early signals and to use the granularity of the research to paint a fuller picture and get more data points.

The causal discovery approach can broaden understanding of organizational safety issues. In this thesis there was the example of "blast furnace gas leak" causing "overflowing contents in waste containers". Assuming this is a valid and recurring phenomenon, the company could proactively change its processes to add additional waste disposal when this type of incident happens.

5.3 Limitations and challenges

The research problem started from the practical question of whether modern natural language processing methods could be utilized for underutilized occupational safety data. Special interest was expressed in finding out the root causes of the incidents. At the early stages of this study, the research question was not yet narrowed down to a single methodological framework. As a result, the study became relatively broad and ambitious, combining unsupervised topic modeling, an LLM-assisted semantic topic structuring experiment, and causal discovery. In retrospect, the study involved a steep methodological learning curve with a rich combination of methods that could not be fully fleshed out in one thesis. More tightly defined research questions with simpler methodology would have improved the methodological coherence and allowed deeper research on fewer components.

This challenge was visible in the early stages of the study, when several alternative NLP approaches were considered before the focus was finally placed on the causal discovery direction. Initial experiments included supervised prediction of accident severity from accident reports with FinBERT, using absence leave as a proxy target (Virtanen et al., 2020). The outcomes were then examined with SHAP to explain the prediction reasoning in an attempt to find out causes for the accident (Lundberg & Lee, 2017). This experiment revealed that the model relied on injury-

related terminology for the prediction task, which limited the model's usefulness for deeper causality approaches. Another attempt was to predict ESAW classifications from the accident text (Eurostat, 2024) using FinBERT as the classification model. Although variants such as masking injury-related terminology before embedding were considered, ultimately the overall classification-based approaches were not aligned with the original research task of being a logical step towards discovering root causes from the data. The exploratory phase was nevertheless impactful, as the ESAW classification task was implemented as part of TAVA's safety system as a helper system for filling incident reports, and the experiments clarified the intuition that classification-based approaches were insufficient for addressing the more open-ended task of root cause analysis.

The shift toward topic modeling was motivated by prior literature on the topic, in which topic modeling was used as part of finding effective COVID treatment options (Kokatnoor & Dr. K, 2022). This new direction was ultimately more compatible with the original research task of finding root causes for accidents. While the study was underway, the field of AI saw rapid development with the release of GPT-3 and the rise of large language models (LLMs) and decoder-only models. This rapid advancement made it difficult to form a stable methodological baseline, creating tension between methodological consistency and the desire to use newer approaches that would improve semantic interpretability.

An obvious gap in the data for modeling the workplace is that there are no starting conditions and no context describing the workplace. The data are mainly limited to actions in reaction to incidents. Possible proactive measures taken in the company that could provide context about routine and previous actions are unknown.

The main problem with topic modeling was that there was no direction or instruction regarding the criteria by which aspects of the input text were the dominant

features on which the clustering was supposed to be based. A more refined model is necessary than the current approach.

An obvious problem with topic modeling is its lack of understanding of how the data should be grouped. Similarly to how severity was often explained by injury words, as the explainability experiment with SHAP showed, the incidents are likely grouped into topics based on similar superficial aspects. For root cause analysis, this type of categorization misses deeper semantic meaning, since a knee injury might be caused by several different reasons. This highlights a fundamental problem in BERTopic, where the model groups words how frequently they appear together, rather than what they actually mean in a causal context. As noted by Dieng et al. (2020), mere lexical co-occurrence is insufficient for capturing complex meaning structures. Instead, models must more effectively leverage the underlying embedding space to understand the deeper functional relationships between terms. Then again, an injury to the eyes and the preventive power of safety glasses provide an actual causal explanation.

5.4 Future research

While the present study focused on semantically structuring incidents into topics and exploring their temporal dependencies, a natural next step would be to treat corrective actions as potential intervention variables. The Granger causality analysis used in this study can identify temporal predictability between topic time series, but it does not by itself establish whether a corrective action has a true causal effect on future incident occurrence. The observed associations may also arise from common causes, seasonality, reporting intensity, or other operational factors.

To move from temporal association towards intervention analysis, a more explicit causal framework would be necessary. One possible approach for this would be to use directed acyclic graphs (DAGs) (Pearl, 2009) to represent the assumed

causal relations between incident types, corrective actions, confounding variables, and future consequences. The idea is to observe what happened and approximate the counterfactual scenario using observational comparison groups.

For example, repeated incidents related to "slipping on ice" could be studied by comparing situations in which gritting was carried out with situations in which it was not. If the frequency of slipping incidents decreased in the gritting group compared to the non-gritting group, the assumption would be that gritting was effective. However, since the control group would be formed from observational data using time and location rather than a randomized experiment, the non-treated group would only be an approximation of the counterfactual scenario rather than a true experimental control group.

This means that possible confounding variables have to be modeled carefully. In the slipping incident, these could be season, weather conditions, time of day, or other contextual information from other reports. A DAG-based causal model could help identify which variables should be controlled for to estimate the effect of the treatment more credibly.

Another approach to analyzing trends would be to utilize hazard incident data as an early warning signal that indicates how specific trends are forming. In practice, the hazard and accident data would need to be clustered together with topic modeling and then aggregated temporally into weekly or monthly bins. Following this, certain accidents could be associated directly with certain hazard data. This is advantageous since hazard data are more numerous than accident data. This would also make it possible to attribute severity to hazards based on the recorded consequences of the accidents.

Further insight could be provided by closely examining the documents located near the topic centroids. These documents could be analyzed with a masking-based approach, for example with a leave-one-out approach by removing individual words

or phrases and comparing the distance of the masked and unmasked versions using cosine similarity. Such analysis could help identify which parts of the text most strongly determine topic membership. This approach follows strategies described by Li et al. (2015) who used similar methods to visualize how individual units contribute to the final composed meaning in neural NLP models.

A repeating motif in this research has been the idea of optimizing not only for semantic similarity but also for causal relevance. Future work could investigate contrastive learning as part of constructing the embeddings, a technique used in causal representation learning (Schölkopf et al., 2021) to move beyond statistical associations. In practice, the embeddings could be created directly from the corrective actions, which would effectively group semantically similar actions into clusters. The goal is to create embeddings that capture the causal relationship between corrective actions (causes) and accidents (effects). After the causal embedder has been trained, it can be used to generate causal vectors.

The embedder could then be fine-tuned to learn these causal embeddings, using the actual corrective action as a positive signal and other unrelated actions as negative signals. Causal embeddings could then be generated for each corrective action in the corpus, after which the standard topic-modeling process would continue.

The study employed only the standard PCMCI methodology, but alternative approaches such as LPCMCI might be better suited to occupational safety data (Reiser, 2022). LPCMCI has the benefit of allowing latent confounders, as could be the case with real-world occupational safety data. In addition to partial correlation, future work could evaluate alternative independence tests such as GPDC or CMIknn, in order to better account for underlying nonlinear dependencies and non-Gaussian distributions.

6 Conclusions

The goal of this study was to build and evaluate a model to identify potential causes of workplace accidents using incident reports as corpora and topic modeling and causal discovery as the chosen methods. Specifically, the aim was to examine whether incident reports written in Finnish can be transformed into meaningful topics and further into aggregated time series that allow the exploration of temporal relationships between them. In the course of the study, the incident data was thoroughly examined, the most relevant parts of the corpus were selected for the analysis, and the data was preprocessed to support both topic modeling and causal analysis.

The results show that the Finnish occupational safety reports used in this study contain enough recurring semantic structure to support automated topic modeling and exploratory temporal analysis. At the same time, the study indicates that moving from semantically coherent topic clusters to reliable causal interpretation remains a substantially more difficult task, especially because of sparse time series data and the limited causal specificity of the topic representations.

With regard to **Research Question I**, the study showed that topic modeling can be used to identify latent topics from workplace safety reports. A robust topic model was built that correctly classified around 60% of incident documents according to LLM-based evaluation. This result suggests that the reports contain enough recurring semantic regularities for topic modeling to produce interpretable groups.

At the same time, the tried approach is limited to general incident categories. The topics are not focused on causally relevant aspects of the incidents.

With regard to **Research Question II**, after classifying the incident reports into topics, for each respective topic, an aggregated time series was formed. These time series were compared to each other with a time lag on the other series to find trending topics following another trending topic. The Granger causality and the PCMCI causality tests provided preliminary results with some links found, which indicated feasibility of the approach with further development of the methodology. This means that temporal dependencies between topic series could be explored, but the evidence was not yet strong enough to support firm causal conclusions in a practical sense.

The main limitations of the study are related to both the topic modeling and the causal discovery. In the topic modeling phase, the generated clusters were not fully representative of a single incident type. The clusters could be formed based on causes, events, consequences or any combination of these. Clear separation of these concerns would need to be established. In the causality section, the biggest problem was the sparsity of the time-aggregated data, which limited the possible links and forced the threshold for detection to be more sensitive. This increases the possibility of stochastic inference and false positive signals. These limitations suggest that the main methodological challenge is not only detecting semantic similarity in the reports, but also forming topic representations that are both semantically coherent and causally informative.

Despite these limitations, the study makes a methodological contribution by demonstrating a complete exploratory pipeline from real-world safety reports to topic-based temporal analysis. The results show that incident reports can be transformed into quantifiable textual patterns and that these patterns can be examined further with time series methods. In this sense, the study provides an initial proof of

concept for combining natural language processing with workplace safety analytics.

Further model development should focus on the semantic quality of the text representations. A major next step to the current plain text implementation of topic modeling would be to develop clustering representations that separate causally relevant aspects of the incident reports, such as underlying causes, incident events, consequences, and contextual conditions. A possible approach would be to utilize LLMs for more refined preprocessing before embedding. Alternatively, future causal analysis would benefit from more granular representations of the reports than aggregated topic frequencies alone. DAGs are a promising avenue to explore that brings multiple possibilities for causal discovery. The NLP and causality fields move fast and future work could evaluate newer methods that better accommodate sparse observations, latent variables and the complex structure of workplace incidents.

In conclusion, this study serves best as exploratory research into the possibilities of utilizing natural language processing methods in the context of workplace safety analysis. The approach taken here was to utilize topic modeling and time-frequency series for causal discovery. This research aims to serve as a starting point for future approaches, ultimately laying a foundation for future research with more rigorous methodology.

References

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in knowledge discovery and data mining* (pp. 160–172). Springer Berlin Heidelberg.

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic modeling in embedding spaces (M. Johnson, B. Roark, & A. Nenkova, Eds.). *Transactions of the Association for Computational Linguistics*, *8*, 439–453. https://doi.org/10.1162/tacl_a_00325

Eichler, M. (2008). Causal inference from time series: What can be learned from granger causality? *Proceedings from the 13th International Congress of Logic, Methodology and Philosophy of Science*.

Elangovan, A., Xu, L., Ko, J., Elyasi, M., Liu, L., Bodapati, S. B., & Roth, D. (2025). Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In Y. Yue, A. Garg, N. Peng, F. Sha, & R. Yu (Eds.), *International conference on learning representations* (pp. 54602–54631, Vol. 2025). https://proceedings.iclr.cc/paper_files/paper/2025/file/8798321486948322be2b4d658744ba72-Paper-Conference.pdf

European Parliament and Council of the European Union. (2008, December). Regulation (ec) no 1338/2008 of the european parliament and of the council of

16 december 2008 on community statistics on public health and health and safety at work [CELEX: 32008R1338].

Eurostat. (2024). *Accidents at work statistics* [Accessed: 2025-04-30]. Eurostat. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Accidents_at_work_statistics

Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., & et al. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. https://arxiv.org/abs/2312.11805

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Grootendorst, M. (2024). *The algorithm – bertopic*. Retrieved April 22, 2025, from https://maartengr.github.io/BERTopic/algorithm/algorithm.html

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, *abs/2004.10964*. https://arxiv.org/abs/2004.10964

Heinrich, H. W. (1941). *Industrial accident prevention: A scientific approach* [Digitized by the Digital Library of India. Scanning Centre: C-DAC, Noida. Source Library: Central Library, BITS Pilani. Identifier: dli.ernet.14601]. McGraw-Hill Book Company Inc. https://archive.org/details/dli.ernet.14601

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2023). *Spacy: Industrial-strength natural language processing in python* [Version 3.8]. https://spacy.io

Hoyle, A. M., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J. L., & Resnik, P. (2021). Is automated topic model evaluation broken?: The incoherence of coherence. *CoRR*, *abs/2107.02173*. https://arxiv.org/abs/2107.02173

Kanerva, J., Ginter, F., Chang, L.-H., Rastas, I., Skantsi, V., Kilpeläinen, J., Kupari, H.-M., Piirto, A., Saarni, J., Sevón, M., & Tarkka, O. (2023). Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for Finnish. *Natural Language Engineering*. https://doi.org/10.1017/S1351324923000086

Kokatnoor, S., & Dr. K, B. (2022). Root cause analysis of COVID-19 cases by enhanced text mining process. *International Journal of Electrical and Computer Engineering*, *12*, 1807–1817. https://doi.org/10.11591/ijece.v12i2.pp1807-1817

Lau, J. H., Newman, D., & Baldwin, T. (2014, April). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In S. Wintner, S. Goldwater, & S. Riezler (Eds.), *Proceedings of the 14th conference of the European chapter of the association for computational linguistics* (pp. 530–539). Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1056

Leveson, N. G. (2012, January 13). *Engineering a safer world: Systems thinking applied to safety*. The MIT Press. https://doi.org/10.7551/mitpress/8179.001.0001

Li, J., Chen, X., Hovy, E. H., & Jurafsky, D. (2015). Visualizing and understanding neural models in NLP. *CoRR*, *abs/1506.01066*. http://arxiv.org/abs/1506.01066

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Madrid-Garcia, A., Freites-Nunez, D., Merino-Barbancho, B., Pérez Sancristobal, I., & Rodriguez-Rodriguez, L. (2024). Mapping two decades of research in rheumatology-specific journals: A topic modeling analysis with bertopic. *Therapeutic Advances in Musculoskeletal Disease*, *16*, 1759720X241308037.

McInnes, L., Healy, J., & Melville, J. (2020, September 18). UMAP: Uniform manifold approximation and projection for dimension reduction. https://doi.org/10.48550/arXiv.1802.03426

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2018). Mixed precision training. https://arxiv.org/abs/1710.03740

Mimino666 & Nakatani, S. (2021). *Langdetect: Port of nakatani shuyo's language-detection library to python* [Version 1.0.8]. https://github.com/Mimino666/langdetect

Mteb leaderboard. (2025). Hugging Face. Retrieved April 28, 2025, from https://huggingface.co/spaces/mteb/leaderboard

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). Mteb: Massive text embedding benchmark. https://arxiv.org/abs/2210.07316

Nedungadi, P., Veena, G., Tang, K.-Y., Menon, R. R. K., & Raman, R. (2025). Ai techniques and applications for online social networks and media: Insights from bertopic modeling. *IEEE Access*, *13*, 37389–37407. https://doi.org/10.1109/ACCESS.2025.3543795

Nguyen, M. V., Lai, V. D., Veyseh, A. P. B., & Nguyen, T. H. (2021). Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. *CoRR*, *abs/2101.03289*. https://arxiv.org/abs/2101.03289

Ohno, T. (1988). *Toyota production system: Beyond large-scale production*. Productivity Press.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, *187*, 253–318.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Raman, R., Pattnaik, D., Hughes, L., & Nedungadi, P. (2024). Unveiling the dynamics of ai applications: A review of reviews using scientometrics and bertopic modeling. *Journal of Innovation & Knowledge*, *9*(3), 100517.

Ramos, L. P. (2026). *Difflib | python standard library* [Real Python, updated April 2, 2026, accessed April 25, 2026]. https://realpython.com/ref/stdlib/difflib/

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Reiser, C. (2022). Causal discovery for time series with latent confounders. https://arxiv.org/abs/2209.03427

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Rooney, J., & Hauvel, L. (2004). Root cause analysis for beginners. *Quality Progress*, *37*, 45–53.

Runge, J. (2024). *Tigramite documentation*. Retrieved May 3, 2026, from https://jakobrunge.github.io/tigramite/

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11), eaau4996. https://doi.org/10.1126/sciadv.aau4996

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Towards causal representation learning. *CoRR*, *abs/2102.11107*. https://arxiv.org/abs/2102.11107

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological methods*, *9*(3), 386.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-b., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., ... Hussenot, L. (2025, March 25). Gemma 3 technical report. https://doi.org/10.48550/arXiv.2503.19786

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2020). Multilingual is not enough: Bert for finnish. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3717–3725.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual e5 text embeddings: A technical report. https://arxiv.org/abs/2402.05672

Wang, Z., Pang, Y., & Lin, Y. (2023). Large language models are zero-shot text classifiers. https://arxiv.org/abs/2312.01044

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*(7), 557–585.