# **Natural language processing for climate**

**Domain expert**

Başak Taraktaş, Boğaziçi University, basak.taraktas@boun.edu.tr

**Writer**

Aybike Şahinoğlu

This study aims to create a corpus from extreme heat and heat wave-related documents gathered from the Prevention Web database and examine their impacts and available mitigation or adaptation strategies using n-grams, TF-IDF, and BERTopic. The analyses are conducted for every category created according to the document content. N-grams and TF-IDF were used the understand context while BERTopic was used to identify main topics in each category. Coherence scores and domain expertise were used to select the most appropriate topics to represent the specific category. The results of the analyses failed to provide meaningful and scientifically relevant results. However, the study still opens an important path for the establishment of a specialized corpus.

*Keywords: heat waves, specialized corpus, Prevention Web*

# **Outline**

**1. Introduction**

1.1. Problem statement

1.2. Research objective

**2. Literature**

2.1. Heat wave and heat vulnerability

2.2. Specialized corpus

2.3. Modeling and evaluation

2.4. How can this research contribute to the literature?

**3. Data and methods**

3.1. Data

3.2. Methods

**4. Results**

4.1. Health and health facilities

4.1.1. Identifying the context

4.1.1.1. TF-IDF

4.1.1.2. N-grams

4.1.2. Identifying the topics

4.1.2.1. Topics and coherence scores

4.1.3. Discussion

4.2. Socioeconomic impacts and resilience

4.3. Disaster risk management

4.4. Risk identification and assessment

4.5. Urban risk and planning

**5. Conclusion**

**6. Limitations**

**7. Annex I**

**8. Annex II**

**9. References**

# **Introduction**

## Problem statement

As the increasing heat patterns turn into hazard and disaster events, actions or policies are needed to protect lives, livelihoods, and assets. However, the public and policymakers often neglect heat waves and ways to prevent their socio-economic-ecological-demographic effects. Therefore, there is a need for more research and “documentation” (ENBEL, 2023), regarding the actions and harms to prevent further damage. So, the visibility of the impacts of heat waves and ways to mitigate as well as adapt must be enhanced to increase awareness among decision-makers, provide guidance, and propose policy solutions.

## Research objective

This study will first establish a corpus from the documents related to extreme heat and heat waves. Documents will be gathered from the Prevention Web (UNDRR) database. Then, the **BERTopic** technique will be utilized on this corpus to identify the main topics that would give ideas about the impacts of heat waves alongside actions for mitigation & adaptation observed in the corpus. The study will use the **coherence score** to evaluate the relevance of the topics.

The main research questions that will guide this objective are:

* How to establish a corpus using documents aiming to mitigate the impacts of heat waves?
* How to identify the topics/characteristics from the documents?

This is an initial step in the establishment of a specialized corpus. As an initial step, it will focus on identifying recurrent issues, making inferences based on them, and providing a path for the further development of this research to become an officially recognized corpus. Therefore, the main focus will be on identifying the context(s) in which the tokens exist and the main topics via topic modeling.

# **Literature**

This section gathers information from the literature regarding the main themes of this research. The last part provides some insights regarding the potential benefits of this study.

### Heat wave and heat vulnerability

Human-related activities have been causing disruptions in the “atmosphere, ocean, cryosphere, and biosphere” (IPCC, 2023, p.5) resulting in disruptions in weather events most importantly in extremely hot days with increased duration and extent accompanied by rising land surface temperature, decrease of vegetation, and diminishing water sources.

Heat waves have various impacts on the environment (e.g. droughts, wildfires, etc.), infrastructures (e.g. blackouts) (Stone et al., 2023), health (exhaustion, fatigue, skin irritation, etc.) (WHO, 2004), economy (e.g. damage to labor productivity and GDP losses) (García-León et al., 2021), education (e.g. lower exam scores) (Srivastava et al., 2017), and many more areas. The intensification of these impacts in terms of harm, damage, duration, and extent increases the possibility of heat waves turning into hazards and eventually disasters. Within this context, vulnerability becomes an important concept to identify. It indicates the groups that are more susceptible to experiencing harm and loss considering demographic, socioeconomic, health, and infrastructural factors (Li et al., 2022) experienced by different communities.

Interventions to mitigate the impacts of heat waves, adapt the environment and populations against heating trends, and increase their resilience vary according to many dimensions such as time, location, and target groups.

Heatwave mitigation refers to the actions and strategies aimed at reducing the impacts of heat waves on human health, infrastructure, ecosystems, and the economy. It may involve interventions such as implementing vegetation on roofs (Tan et al., 2023), establishing cooling centers (Meade et al., 2023), etc. Heatwave adaptation involves implementing measures to adjust and respond to the challenges posed by heat waves. It may involve raising awareness regarding the dangers of heat waves, strengthening the infrastructures (Kiarsi et al., 2023), etc.

### Specialized corpus

Within the social and public policy literature, specialized corpora have been established for many tasks such as identifying patterns or trends by gathering policy documents (Ridgway, 2023), developing NLP methods to improve policy analysis (Safaei & Longo, 2024) , and mining public opinions (Zha et al., 2023).

*DCEP: Digital Corpus of the European Parliament* is a good example of a specialized corpus involving unstructured data for policy analysis. Within the climate change mitigation and adaptation literature, there are specialized corpora for processing climate program/project documents to evaluate their alignment with policies (Jin et al., 2023), conducting discourse analysis, examining climate change trends (Volkanovska et al., 2023), and understanding the effects of climate change in different locations (Mallick et al., 2024).

*Climate Policy Radar* is an important example that defines and involves climate policy-related documents across different languages. Within hazard and disaster literature, there are studies investigating the perceptions and sentiments regarding extreme heat (Zhu et al., 2024), responses against heat waves (Zander et al., 2023), detecting air quality (Văduva et al., 2023), and assessing the impact of heat-related hazards (Sodoge et al., 2023) using different sources such as social media, news, and archives.

Using social media to detect sentiments, opinions, and responses against heat waves is the most common research design to understand the impact of heat waves. There are also literature reviews to understand the drivers behind what makes certain groups vulnerable to increasing temperatures (Li et al., 2022) and studies that quantify vulnerability (Bayomi & Fernandez, 2023). However, there is no specialized corpus like DCEP or Climate Policy Radar involving heatwave impacts, mitigation & adaptation interventions, policies, and plans to inform decision-makers.

### Modeling and evaluation

The studies in the literature generally use the Latent Dirichlet Algorithm (LDA) model to identify topics, issues, and impacts. LDA is an “unsupervised learning algorithm” (Supiadin & Laksito, 2023, p.3331) that groups relevant words under each topic. Since LDA is a statistical model, it cannot account for the semantic relations between words (Zengul et al., 2023). It is possible to calibrate LDA models with domain knowledge, however, it is time and effort-consuming. Instead, BERTopic can better identify topics in small and unstructured datasets while accounting for contextual relations.

The coherence score is generally used for evaluating models by quantifying the closeness and interpretability of words within a topic. It is used to assess the appropriate number of topics that cover the features of interest.

### How can this research contribute to the literature?

This study can contribute to the heat wave and heat vulnerability literature by creating a specialized corpus that includes textual documents from multiple sources around the world that accounts for different impacts of heat waves and different groups who are / might be susceptible to experiencing greater harm.

This corpus can become an important international resource for studying the impacts of heat waves on different geographical regions, demographics, and socio-economic groups. Policymakers can use insights from this corpus to develop targeted policies and interventions to mitigate the impacts of heat waves, especially for vulnerable populations.

This corpus can help better understand which groups are most at risk and guide the allocation of resources and the design of effective adaptation strategies. For example, public health officials can use the corpus to identify at-risk populations and develop targeted health interventions and communication strategies. This could include heat wave preparedness plans, early warning systems, and outreach campaigns tailored to specific local communities. Educators, media professionals, and even advocacy groups can use the corpus to raise awareness about the impacts of heat waves. It can help disseminate information about heat wave risks and protective measures tailored to different audiences. Moreover, this corpus can encourage further research to include multimodal data (Volkanovska et al., 2023). From the computational social science perspective, this corpus can be used for topic modeling to identify themes and patterns, geospatial analysis the locate impacts and interventions, network analysis to identify the relations between stakeholders in implementing interventions, and perhaps predictive modeling to create optimized interventions.

# **Data and methods**

This section introduces and explains the database, the structure of the data, and the methods used.

## Data

The data is gathered from the Prevention Web “Knowledge Base” (Prevention Web, n.d.). The data is gathered from the “Research Briefs” and “Documents & Publications “sections, filtered by hazard (heat wave). Documents in this database are gathered from different sources (government archives, journals, etc.) worldwide.

The documents are in different formats (reports, research articles, discussion papers, policy briefs, policy recommendations, guidelines, reviews, technical notes, briefings, etc.). Therefore, they are unstructured. Meaning that their format (text font & alignment, existence of pictures & alignment, etc.) differ. They are all in PDF format and in English[[1]](#footnote-1). However, some of the documents in this dataset are corrupt, cannot be downloaded from the website, cannot be found on the web, or are unrelated[[2]](#footnote-2).

The PDF documents were downloaded from the Prevention Web and stored on the local machine. There are 528 documents in total ready for this research. Moreover, the documents cover the years of 2008 and 2024.

Within the Prevention Web database, each document is categorized under at least one “theme” (Prevention, Web, n.d.) (e.g. governance). The downloaded documents were categorized and analyzed according to these themes for efficiency in processing and providing detailed results. Each document’s category is based on the domain knowledge and the database’s decision.

The categories involve **Health and health facilities**[[3]](#footnote-3) (108 documents), **Socioeconomic impacts and resilience**[[4]](#footnote-4) (105 documents), **Urban risk and planning**[[5]](#footnote-5) (79 documents), **Risk identification and assessment**[[6]](#footnote-6) (115 documents), and **Disaster risk management** [[7]](#footnote-7)(121 documents).

## Methods

This paper provides an initial step in the establishment of a specialized corpus. As an initial step, it will focus on identifying the main issues, analyzing them, and providing a path for the further development of this research to make it an officially recognized corpus. Therefore, the main methods of this research are identifying the **context(s)** and the main **topics**.

Identification of the context(s) is conducted with **n-grams** as well as **TF-IDF** scores and identification of main topics is conducted with **BERTopic**.

N-grams is a method for identifying word-co-occurrences in a corpus. N-grams models can capture the co-occurrences of n number of words. For this study, the trigrams (three-word occurrences) was used to capture the context of each category.

TF-IDF (Term Frequency-Inverse Document Frequency) is a method that measures the importance of words considering how often a certain word appears in a document (TF) and its prevalence in every document.

N-grams and TF-IDF together can be used to generate insights about the context by looking at the important words operationalized by TF-IDF scores and frequent word co-occurrences based on n-grams.

BERTopic is an advanced topic-modeling method that generates automated topics from the corpus. It utilizes the BERT (Bidirectional Encoder Representations from Transformers) language model, which is trained on diverse data, to create embeddings for each document. It also utilizes c-TF-IDF (class-based TF-IDF) to examine the meaningful relations between these embeddings. Thus, BERTopic can generate topics while accounting for the semantic and contextual relations between tokens.

Identifying the main topics involves examining the results of the BERTopic and their coherence. The topic model involves,

* min\_topic\_size[[8]](#footnote-8)
* top\_n\_words[[9]](#footnote-9)
* n\_gram\_range[[10]](#footnote-10)
* calculate\_probabilities[[11]](#footnote-11)

as main parameters. The reason for selecting these parameters is to create the most interpretable results both for a domain expert and researchers from different fields. The parameter adjustment is a trial-error process based on the domain expertise and coherence scores.

**“min\_topic\_size”** (Grootendorst, 2024) is important for determining how the BERTopic model should construct the context for each topic. Ideally, at least ten documents are sufficient for creating understandable topics. Topics were constructed with minimum five topics for *Health and health facilities, Disaster risk management,* and *Urban risk and planning* categories. On the other hand, topics were constructed with minimum two documents for *Risk identification and assessment,* and *Socioeconomic impacts and resilience* categories. This is because the number of documents per category is quite low. Trial and error with changing the min\_topic\_size showed that a smaller number of documents are needed to create understandable topics in categories that involve documents that are not very related to each other.

**“top\_n\_words”** (Grootendorst, 2024) is important for incorporating the appropriate words to represent a topic. Each topic is represented by five words for *Health and health facilities* and *Risk identification and assessment* categories. On the other hand, each topic is represented by six words for *Disaster risk management, Urban risk and planning,* and *Socioeconomic impacts and resilience* categories. There is no specific reasoning behind the selection of top\_n\_words other than trial and error.

**“n\_grams\_range”** (Grootendorst, 2024) is important for considering unigrams and bigrams in accounting for the context. n\_gram\_range is selected as (1, 2) for every category.

**“calculate\_probabilities”** (Grootendorst, 2024) is important for calculating the probabilities of each topic for every document. It is set to “True” for every category.

The evaluation of the context is based on the domain knowledge. The evaluation of the topics, on the other hand, is based both on coherence scores and domain knowledge.

The analyses are conducted separately according to each category. However, they all share the same workflow[[12]](#footnote-12) :

* Generating text data from PDFs
* Preprocessing & word tokenization
* Defining token dictionary
* Defining bag of words (BoW) representation
* TF-IDF modeling
* N-grams modeling
* BERTopic modeling
* Calculating coherence scores for each topic

# **Results**

This section explains the results of topic modeling and evaluation. The results will be discussed according to each category.

## Health and health facilities

This category involves 108 documents and 27005 unique word tokens.

### Identifying the context

#### TF-IDF score

After several trials, it was decided that tokens that have a score above 0.05 should be considered appropriate for analyzing the context for every category. The reason for selecting this number as the threshold is because the words with TF-IDF scores 0.05 and over are more meaningful according to the domain knowledge. Unfortunately, due to the problems with stop words mentioned in the workflow, there are non-relevant words that have a score above 0.05. Still, it is possible to observe some important occurrences[[13]](#footnote-13) that overlap with the domain knowledge.

As can be observed from the TF-IDF table[[14]](#footnote-14), this category is mainly about the impacts of heat waves and extreme heat on pulmonary health, pregnancy, internal organs, neonatal development, morbidity, and mortality. Moreover, women, children, and the elderly seem to be the most identified vulnerable groups.

#### N-grams

After several trials, it was decided that trigram (with three tokens) n-grams provide the most meaningful results. Unfortunately, problems with the stop words persist in the n-grams results. Therefore, a domain knowledge filter was needed. Moreover, due to the sheer number of n-grams, only the important results will be discussed. You can refer to the code to access the entirety of n-grams.

Alongside the mentioned groups above, it is possible to observe connotations that highlight the importance of **socioeconomic status** ('disadvantage', 'neighbourhood', educational', 'attainment’) as functions of **vulnerability**. As impacts, some connotations identify different **health aspects** that do not exist in TF-IDF scores ('bacteriological', 'histological', 'confirmation', 'cardiovascular', 'disorder', 'originate').

### Identifying the topics

#### Topics and coherence scores

For each category, coherence scores and domain knowledge were used together to identify the most meaningful topic(s).

The BERTopic model for this category involves “min\_topic\_size=5” and “top\_n\_words=5” for the most meaningful results.

The BERTopic model created three topics[[15]](#footnote-15) represented by the word lists below.

|  |  |
| --- | --- |
| **Topic** | **Representative word list** |
| -1 | [wave, child, physical, skin, compound] |
| 0 | [mortality, city, air, day, population] |
| 1 | [exposure, birth, child, pregnancy, maternal] |

The coherence scores[[16]](#footnote-16) have generated these results for each topic.

|  |  |
| --- | --- |
| **Topic** | **Coherence score** |
| -1 | 0.51 |
| 0 | 0.49 |
| 1 | 0.76 |

One topic with a coherence score below 0.5[[17]](#footnote-17) was disregarded.

The **first** topic involves the words below with their probabilities,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| child | 0.0242 |
| physical | 0.0229 |
| skin | 0.0206 |
| disease | 0.0182 |
| infection | 0.0164 |

The coherence score for this topic is 0.51. Deriving from domain knowledge, extended exposure to heatwaves results in skin irritation/diseases and increases the possibility of infectious disease occurrence, spread, and extent.

The **third** topic involves the words below with their probabilities,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| exposure | 0.0369 |
| birth | 0.0332 |
| child | 0.0288 |
| pregnancy | 0.0282 |
| maternal | 0.0236 |

The coherence score for this topic is 0.76. Deriving from domain knowledge, pregnant women, children, newborns, and infants are one of the most vulnerable groups against exposure to heat waves.

### Discussion

The topics, their coherence score, and the patterns identified seem to perpetuate each other. It seems that the BERTopic mainly generates topics that are measured as important by the TF-IDF model. However, not every relation (even though they can be considered meaningful) captured by the n-grams model is identified as a main topic.

## Socioeconomic impacts and resilience

This category involves 105 documents and 33894 unique tokens.

### Identifying the context

#### TF-IDF score

As can be observed from the TF-IDF table, this category is mainly about the impacts of heat waves on work conditions (especially outside work such as agriculture), livelihoods, productivity, and health. In this category, it is possible to observe emphasis on the health impacts of heat waves. This could be because socioeconomic and health situations are closely intertwined and the documents in this category are mainly about the impacts of heat waves on outside workers.

According to these results, the most vulnerable groups identified in this category are outside workers, poor individuals and households, the elderly, women, and the young.

#### N-grams

Alongside the connotations that perpetuate the relationships mentioned above, there are emphases on the **educational impacts of heat waves** ('low', 'test', 'score') and the **conditions of students** ('school', 'locate', 'hot'). Moreover, there are emphases regarding **food safety** ('food', 'security', 'alert'), **economic volatilies** ('baseline', 'inﬂation', 'volatility') ('growth', 'shock', 'wealth'), **inequalities** ('inequality', 'consumption', 'inequality') ('black', 'hispanic', 'population'), and **violence** ('crime', 'violent', 'sexual')

### Identifying the topics

#### Topics and coherence scores

The BERTopic model for this category involves “min\_topic\_size=2” and “top\_n\_words=6” for the most meaningful results.

The BERTopic model created fourteen topics represented by the word lists below[[18]](#footnote-18).

|  |  |
| --- | --- |
| **Topic** | **Representative word list** |
| -1 | [exposure, population, vulnerability, air] |
| 0 | [disaster, country, development, loss, people] |
| 1 | [worker, work, stress, occupational, labour] |
| 2 | [ination, shock, region, future, hws, population] |
| 3 | [emission, energy, trade, technology, cool] |
| 4 | [unit, insurance, tdcj, texas, hate, tweet] |
| 5 | [drought, vegetation, ozone, global, use, lst] |
| 6 | [migration, tourism, australia, australian] |
| 7 | [stratum, day, mean, electricity, household] |
| 8 | [event, return, firm, excess return, excess, day] |
| 9 | [location, threshold, fan, neighbourhood] |
| 10 | [damage, loss, drought, cost, germany, wood] |
| 11 | [cool, roof, cool roof, access, energy, wave] |
| 12 | [mortality, city, tract, uhi, lst, age] |

The coherence scores have generated these results for each topic.

|  |  |
| --- | --- |
| **Topic** | **Coherence score** |
| -1 | 0.59 |
| 0 | 0.49 |
| 1 | 0.83 |
| 2 | 0.36 |
| 3 | 0.62 |
| 4 | 0.27 |
| 5 | 0.41 |
| 6 | 0.58 |
| 7 | 0.61 |
| 8 | 0.65 |
| 9 | 0.59 |
| 10 | 0.62 |
| 11 | 0.67 |
| 12 | 0.70 |

Four topics with coherence scores below 0.5 were disregarded.

When all the topics were evaluated with domain knowledge, it was decided that only four topics involved the most interpretable and meaningful results.

The **third** topic involves the words below with their probabilities,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| worker | 0.0393 |
| work | 0.0345 |
| stress | 0.0277 |
| occupational | 0.0153 |
| labor | 0.0137 |
| productivity | 0.0136 |

The coherence score for this topic is 0.83. Deriving from domain knowledge, increased exposure of outdoor workers to heat decreases labor productivity and results in serious physical strain as well as stress.

The **eighth** topic involves,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| migration | 0.0417 |
| tourism | 0.0331 |
| australia | 0.0261 |
| australian | 0.0167 |
| migrant | 0.0142 |
| study | 0.0109 |

The coherence score for this topic is 0.58. Deriving from domain knowledge, increasing the extent and duration of heat days increases the possibility of in and out-migration. Moreover, as hazards escalate to disasters, the inhabitable circumstances increase the probability of forced and mass migrations. Also, as economic activity slows down, the tourism sector loses its productivity and consumers.

The **twelfth** topic involves,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| damage | 0.0329 |
| loss | 0.0286 |
| drought | 0.0220 |
| cost | 0.0214 |
| germany | 0.0198 |
| wood | 0.0170 |

The coherence score for this topic is 0.62. Deriving from domain knowledge, increasing the extent and duration of heat days increases the possibility of drought and consequent vegetation loss. This situation also results in economic losses.

The **fourteenth** topic involves,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| mortality | 0.0339 |
| city | 0.0265 |
| tract | 0.0252 |
| uhi | 0.0244 |
| lst | 0.0182 |
| age | 0.0175 |

The coherence score for this topic is 0.70. Deriving from domain knowledge, increasing the extent and duration of extreme heat days increases the land surface temperature (lst) and the possibility of the urban heat island effect (uhi), especially in urban areas. This results in excess mortality especially among very young and old populations.

### Discussion

This category involves diverse subthemes that may be difficult to aggregate to create meaningful topics. That is why, the minimum number of documents to be observed by BERTopic was limited to two. The topics seem to perpetuate the results gathered from TF-IDF and n-grams analyses, however, the relations between poverty and heat waves seem to be missing in the topic model results. Of course, this depends on the number of documents that investigate this topic.

Even though the coherence scores of some topics are above 0.5, they do not make much sense when the topic words are considered from the domain perspective.

## Disaster risk management

This category involves 121 documents and 35629 unique tokens.

### Identifying the context

#### TF-IDF score

The results in the TF-IDF table seem to be sparse. Impacts such as overheating (e.g. fires) seem to be the risks related to heat waves and extreme heat according to this category. As for the management part, it is possible to observe governance-related means (e.g. policy and warning) and a mention of technology. There seems to be no mention of vulnerable groups.

#### N-grams

Alongside the connotations that perpetuate the relationships mentioned above, there are emphases on **vulnerability** ('predict', 'neighborhood', 'vulnerability') ('government', 'organization', 'vulnerable'), increase in emissions ('energy', 'demand', 'emission'), **food security & malnutrition** ('malnutrition', 'rate', 'increase'), and **industrial activities** ('oil', 'palm', 'plantation') as risks. **Community engagement for enhancing resilience** ('community', 'science', 'involve'), **residential cooling** ('residential', 'heat', 'cool'), **the use of information technology** ('information', 'technology', 'enable'**), the leveraging of science** ('integrate', 'available', 'science'), **collaboration** ('coordination', 'collaboration', 'agency'), **sustainable agriculture** ('sustainable', 'agricultural', 'practice'), **conservation** ('protection', 'natural', 'ecosystems'), and **inclusion** ('indigenous', 'crop', 'traditional') as management strategies.

### Identifying the topics

#### Topics and coherence scores

The BERTopic model for this category involves “min\_topic\_size=5” and “top\_n\_words=6” for the most meaningful results.

The BERTopic model created seven topics represented by the word lists below[[19]](#footnote-19).

|  |  |
| --- | --- |
| **Topic** | **Representative word list** |
| -1 | [water, development, include, need, system, area] |
| 0 | [event, weather, europe, system, warm, change] |
| 1 | [disaster, information, thailand, flood] |
| 2 | [food, water, country, system, crop] |
| 3 | [cool, energy, demand, thousand, ac, air] |
| 4 | [ahmedabad, cool, disaster, people, karachi] |
| 5 | [event, attribution, event attribution, wave] |

The coherence scores have generated these results for each topic.

|  |  |
| --- | --- |
| **Topic** | **Coherence score** |
| -1 | 0.45 |
| 0 | 0.54 |
| 1 | 0.54 |
| 2 | 0.60 |
| 3 | 0.68 |
| 4 | 0.50 |
| 5 | 0.48 |

Two topics with coherence scores below 0.5 were disregarded.

When all the topics were evaluated with domain knowledge, it was decided that only two topics involved the most interpretable and meaningful results.

The **fourth** topic involves,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| food | 0.0276 |
| water | 0.0181 |
| country | 0.0137 |
| system | 0.0128 |
| crop | 0.0113 |
| agricultural | 0.0110 |

The coherence score for this topic is 0.60. Deriving from domain knowledge, increasing the extent and duration of heat days depletes the water sources which jeopardizes the agricultural output, crop qualities, and food security.

The **fifth** topic involves,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| cool | 0.0318 |
| energy | 0.0256 |
| demand | 0.0158 |
| thousand | 0.0153 |
| ac | 0.0147 |
| air | 0.0133 |

The coherence score for this topic is 0.68. Deriving from domain knowledge, increasing the extent and duration of heat days increases the demand for cooling (e.g. a/c air conditioning) and energy consumption.

### Discussion

This category also involves diverse subthemes that may be difficult to aggregate to create meaningful topics. Therefore, the topic model results are sparse. The ambiguity in defining the category and selecting the appropriate documents based on it may be one of the reasons for sparsity and inadequate emphasis on mitigation and adaptation actions for managing disaster risks.

## Risk identification and assessment

This category involves 115 documents and 39403 unique tokens.

### Identifying the context

#### TF-IDF score

The results in the TF-IDF table do not provide meaningful interpretations.

#### N-grams

In n-grams results, there are emphases on ('extremely', 'warm', 'summer') and ('concurrent', 'heatwave', 'intensity') as risks that can create harm to lives, livelihoods, and assets. However, there are no meaningful connotations in the model.

### Identifying the topics

#### Topics and coherence scores

The BERTopic model for this category involves “min\_topic\_size=2” and “top\_n\_words=5” for the most meaningful results.

The BERTopic model created fourteen topics represented by the word lists below[[20]](#footnote-20).

|  |  |
| --- | --- |
| **Topic** | **Representative word list** |
| -1 | [water, global, impact, hazard, risk] |
| 0 | [flood, vulnerability, mumbai, population, area] |
| 1 | [ehe, green, water, cool, element] |
| 2 | [change, project, water, precipitation, rise] |
| 3 | [drought, water, event, mean, precipitation] |
| 4 | [australia, queensland, fire, heatwave] |
| 5 | [forecast, mortality, cool, weather, day] |
| 6 | [heatwave, heatwaves, trend, day, wave] |
| 7 | [european, europe, disaster, natural, eu] |
| 8 | [anomaly, precipitation, attribution] |
| 9 | [fishery, chwesl, chwesl event, marine, sea] |
| 10 | [utci, ta, tibet, hw, trend] |
| 11 | [mortality, wave, mortality rate, rate, ewes] |
| 12 | [drought, sector, europe, water, australia] |

The coherence scores have generated these results for each topic.

|  |  |
| --- | --- |
| **Topic** | **Coherence score** |
| -1 | 0.54 |
| 0 | 0.59 |
| 1 | 0.64 |
| 2 | 0.62 |
| 3 | 0.57 |
| 4 | 0.76 |
| 5 | 0.55 |
| 6 | 0.63 |
| 7 | 0.77 |
| 8 | 0.66 |
| 9 | 0.47 |
| 10 | 0.52 |
| 11 | 0.75 |
| 12 | 0.57 |

One topic with a coherence score below 0.5 was disregarded.

When all the topics were evaluated with domain knowledge, it was decided that only one topic involved the most interpretable and meaningful results.

The **sixth** topic involves,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| australia | 0.0250 |
| queensland | 0.0248 |
| fire | 0.0219 |
| heatwave | 0.0174 |
| australian | 0.0148 |

The coherence score for this topic is 0.76. Deriving from domain knowledge, heat waves increase the possibility of fires in urban and rural areas.

### Discussion

So far, this category produced the worst results. This may be due to the ambiguity in defining the category and selecting the appropriate documents based on it. The documents in this category distantly crosscut the other categories, but do not fit any of them. Combined with the high sparsity of the themes, the results from this category fail to establish a coherent context and produce interpretable topics.

## Urban risk and planning

This category involves 79 documents and 26107 unique tokens.

### Identifying the context

#### TF-IDF score

According to the table, indoor and outdoor overheating seems to be one of the most important risks of heat waves in urban areas. Overheating is affected by building features and surface as well as tree canopy that also affect the cooling effect. The results also include one effect of overheating (e.g. crime), means to tackle harm and vulnerabilities (e.g. policy and financing).

#### N-grams

Alongside the connotations that perpetuate the relationships mentioned above, there are emphases on **building modifications** ('green', 'roof', 'control'), **green infrastructure** ('green', 'infrastructure', 'mitigation'), **thermal comfort** ('thermal', 'comfort', 'island'), **greening for cooling** ('mean', 'cool', 'tree'), **cooling techniques** ('green', 'roofs', 'wetland'), **warning systems** ('warning', 'system', 'evacuation'), **increasing community resilience** ('enhancement', 'community', 'resilience'), **increasing institutional capacity** ('institution', 'capacity', 'building'), **increasing governance engagement** ('encouragement', 'government', 'engagement'), and **prioritizing critical infrastructure** ('prioritise', 'critical', 'infrastructure') as mitigation strategies. **Excess mortality** ('summer', 'preventable', 'death'), **urban heat island effect** ('mortality', 'attributable', 'uhi') and **drought** '(hotter', 'drier', 'summer')('demand', 'water', 'increase') as the impacts of extreme heats in urban areas. Moreover, ('age', 'ethnicity', 'commute') ('ethnicity', 'commute', 'employment') ('access', 'healthcare', 'service') as identifiers of vulnerability in urban areas.

### Identifying the topics

#### Topics and coherence scores

The BERTopic model for this category involves “min\_topic\_size=6” and “top\_n\_words=5” for the most meaningful results.

The BERTopic model created six topics represented by the word lists below.

|  |  |
| --- | --- |
| **Topic** | **Representative word list** |
| -1 | [london, area, water, flood, use] |
| 0 | [cool, green, energy, building, roof] |
| 1 | [exposure, population, city, area, future] |
| 2 | [roof, energy, cool, resilience, green] |
| 3 | [city, cool, energy, policy, area] |
| 4 | [heatwave, people, cross, community, heatwaves] |

The coherence scores have generated these results for each topic.

|  |  |
| --- | --- |
| **Topic** | **Coherence score** |
| -1 | 0.52 |
| 0 | 0.73 |
| 1 | 0.54 |
| 2 | 0.66 |
| 3 | 0.50 |
| 4 | 0.69 |

All coherence scores were above 0.5, there were no disregards.

When all the topics were evaluated with domain knowledge, it was decided that only one topic involved the most interpretable and meaningful results.

The **fourth** topic involves,

|  |  |
| --- | --- |
| **Word** | **Probability** |
| roof | 0.0246 |
| energy | 0.0179 |
| cool | 0.0165 |
| resilience | 0.0160 |
| green | 0.0153 |

The coherence score for this topic is 0.66. Deriving from domain knowledge, green roofing is one of the most important strategies to provide cooling and save energy.

### Discussion

The category involves fewer documents. Therefore, the topics are not diverse and it is possible to find more content in n-grams results.

# **Conclusion**

This study provided insights and analyses as an initial step in the establishment of a specialized corpus based on the diverse impacts of heat waves and ways to mitigate & adapt. The results gathered from analyzing the context and main topics of the categories, mainly inform about the impacts of heat waves on different groups and issues. There was not much emphasis on mitigation and adaptation measures. Nonetheless, these findings are important in understanding the pressing issues around the world related to heat waves. For this purpose, n-grams, TF-IDF, and BERTopic models were helpful. In identifying the context, the TF-IDF model was quite useful in identifying the most recurrent themes in the categories and n-grams was useful in detailing the themes and showing other connotations that are important for the domain but unaccounted by TF-IDF. In identifying the topics, BERTopic was detailed and efficient.

It is possible to generate insights below by analyzing n-grams, TF-IDF, and BERTopic results together.

For the *Health and health facilities* category, the documents involved in this category emphasize the impact of extreme heating and heat waves on maternal, neonatal, and child health. Moreover, exposure can also impact skin health and cause diseases and infections. Socioeconomic status, as a function of vulnerability, may be an important reason for higher exposure to heat and related health outcomes. Targeted mitigation and adaptation strategies are not explicitly mentioned in the topic model alongside n-grams and TF-IDF results.

For the *Socioeconomic impacts and resilience* category, the documents involved in this category emphasize the impact of extreme heating and heat waves on educational outcomes, food safety, economic volatilities, worker conditions, labor productivity, migration, tourism, drought, and vegetation loss mainly because of the increasing land surface temperature and urban heat island effect. Targeted mitigation and adaptation strategies are not explicitly mentioned in the topic model alongside n-grams and TF-IDF results.

For the *Disaster risk management* category, the documents involved in this category emphasize the impact of extreme heating and heat waves on food security & malnutrition, water resources, and agricultural production as the main risks. Community engagement & inclusion, residential cooling, the use of information technology, the leveraging of science,collaboration, conservation, and sustainable practices in production are emphasized as management strategies.

For the *Risk identification and assessment* category, the documents involved in it emphasize the impact of extreme heating and heat waves on the increasing possibility of fires.

For the *Urban risk and planning* category, the documents involved in this category emphasize the impact of extreme heating and heat waves on excess mortality and drought mainly as the results of the increasing urban heat island effect. Building modifications, implementing green infrastructure, cooling, implementing warning systems, strengthening community resilience, enhancing institutional capacity, and increasing government engagement are emphasized as the main mitigation strategies.

The results and insights gathered from these categories cannot by any means be generalized as the documents in the Prevention Web database are mainly gathered from English-speaking countries. Meaning that there is an overrepresentation of certain countries. This is a major bias that needs to be addressed to make this corpus useful for the international audience and researchers from diverse backgrounds.

Due to the limitations with data and methods alongside narrow results generated by the analyses, this study failed to answer its research questions and generate meaningful results. Still, as a first step, it opens an important path and provides motivation for redemption and further engagement.

# **Limitations**

The study involves serious limitations with data, methods, and results.

Limitations with data,

* The Prevention Web database heavily involves documents from English-speaking countries.
* Some countries such as Australia and India are over-represented.
* Although categories were based on domain knowledge and Prevention Web’s classification, problems with the number of documents and misclassification resulted in non-meaningful results in some categories.
* The content of the categories may overlap with each other.
* The number of documents for each category is relatively low.

Limitations with methods,

* Stop words continue to occur despite cleaning attempts in the preprocessing step.
* Possible erroneous adjustments with the BERT topic model could have lowered the interpretability fo the topics.
* Running the same code without changing the parameters of the BERT topic provides different topic results.
* There is no conceptualization and operationalization of the biases involved in every step of this research.

Limitations with results,

* In evaluating the context with TF-IDF and n-gram models, only the domain knowledge was relied upon. Therefore, the terms “meaningful” and “interpretable” remain heavily subjective.
* In evaluating the topic models using the coherence score, a mixture of domain knowledge and coherence score was relied upon. The problem with this evaluation is that the topic with the highest coherence score does not mean it is the most meaningful one. In this case, the implications of the coherence score remain ambiguous.
* The insights cannot be generalized because they are confined to the context established by the documents, which is not large enough.
* There is a heavy reliance on domain knowledge rather than numerical results in creating meaningful results and interpretations.
* The research reported only the working and most meaningful results in order not to confound the overall insights and make this study any longer.

# **Annex I**

## Workflow

## Documents are gathered and categorized

## Texts are gathered from PDF documents

The documents are in different formats (reports, research articles, discussion papers, policy briefs, policy recommendations, guidelines, reviews, technical notes, briefings, etc.). Therefore, they are unstructured, which means that their format (text font & alignment, existence of pictures & alignment, etc.) differ. They are all in PDF format and in English. However, some of the documents in this dataset are corrupt, cannot be downloaded from the website, cannot be found on the web, or are unrelated.

“PyMuPDF” library was used to get textual data from the documents.

## Preprocessing is conducted

Preprocessing was the most tedious and problematic part of this study.

First, “WordNet POS” tags for lemmatization were loaded. Then, a preprocessing pipeline was defined.

The preprocessing pipeline involves,

* Removing multiple white spaces and line breaks
* Applying “nltk.word\_tokenize” on clean text
* Loading NLTK stopword list and defining a custom stopword list[[21]](#footnote-21)
* Standardizing the text and removing non-alphabetic or stop words
* Defining lemmatizer
* Conducting POS tagging
* Lemmatizing word-tokens via assigned POS tags

An iteration function was defined afterward. The purpose of the iteration function is to go over each page of each document to confirm that they are in the desired format (PDF), apply the preprocessing pipeline, and store tokens generated for each file in a dictionary. The iteration function involves,

* Creating an empty dictionary to display all documents and tokens together
* Defining a counter to track the number of processed texts
* Defining a for loop to iterate over each file to
* check whether the files are PDFs,
* create a path to the PDF files, extract the content of files,
* apply the preprocessing pipeline,
* generate a unique name for each document,
* assign the preprocessed text content of the current document in the docs dictionary with the document name as the key
* Defining the pdf folder path
* Printing each document's tokens as a list

## Token dictionary is defined

Defining a token dictionary is important for preparing tokens for Bag of Words (BoW) representation and related applications by assigning each token a unique ID and collecting them under a single variable. The gensim library was utilized for this purpose.

## BoW representation is defined

BoW representation is important for converting tokens into numerical vectors, where each dimension corresponds to a unique token in the dictionary. It retrieves the frequency of each token in a document. This step makes the data ready for TF-IDF and n-grams applications.

## TF-IDF model is defined

TF-IDF is a model that captures the importance of tokens by weighing the frequency of a token (TF) against its frequency in the entire corpus (IDF). Tokens that appear frequently in a specific document but rarely in other documents are considered more important. Although TF-IDF is mainly used as input for LDA, it can be useful for providing inputs for n-gram analysis. Moreover, TF-IDF can be useful for identifying stop words.

Identifying stop words with TF-IDF can be tricky. It basically involves defining a probability threshold (0.05) and defining the words below this threshold as stop words. This method was not used fearing that there would not be many words to work with and certain domain-specific words would be removed.

## N-grams model is defined

N-grams is a model that involves a collection of words that appear together according to their frequency of co-occurrence. Therefore, n-grams is important for examining the context in which the tokens appear. For this study, trigram (with 3 words) n-gram was used because it can provide a better idea regarding the context.

## BERTopic model is defined

As a statistical model, LDA fails to identify the contextual relationships between the tokens. It merely identifies topics based on word distributions, which may fail to capture coherent topics generated from small and technical datasets. That is why, BERTopic was used.

BERTopic is a model based on transformer architecture that utilizes embeddings. Within the scope of this study, the BERTopic with "all-MiniLM-L6-v2" sentence transformer was used to create embeddings that help generate meaningful topics.

After adjusting the parameters, topics were displayed. The adjustment step is an iterative process depending on the comparison of domain knowledge with coherence scores.

## Coherence scores are calculated

The coherence score is a measure used to evaluate the degree of similarity between words within the same topic. A higher coherence score indicates that the words in the same topic are more closely related, which indicates that the topic is more meaningful and interpretable. The coherence score is important for evaluating to what degree topics and the words they involve are interpretable by a domain expert and identifying words that are likely to be relevant to the themes present in the corpus.

The “gensim.models.coherencemodel” was used to calculate topic coherence scores.

# **Annex II**

## Health and health facilities

### TF-IDF[[22]](#footnote-22)

|  |  |
| --- | --- |
| Word | TF-IDF score |
| comfort | 0.0659437292376771 |
| birth | 0.10151896794257115 |
| air | 0.0686126892861949 |
| humidity | 0.16607846327843365 |
| age | 0.10464880427795568 |
| pollutant | 0.05711929712000558 |
| pulmonary | 0.058635357055859996 |
| maternal | 0.1610014978287042 |
| postpartum | 0.16601096001522245 |
| pregnant | 0.3287451203735829 |
| woman | 0.20855302428196293 |
| exhaustion | 0.07515562977224091 |
| kidney | 0.06780184116288421 |
| breastfeed | 0.5061129198207569 |
| gerontological | 0.06808475331837915 |
| infant | 0.24569720372504358 |
| neonatal | 0.0948975528923031 |
| newborn | 0.20357830820326486 |
| paediatrics | 0.06159821136974789 |
| placenta | 0.050894577050816216 |
| postnatal | 0.06159821136974789 |
| pregnancy | 0.1562490468497292 |
| preterm | 0.09464793264177107 |
| old | 0.11988596802488666 |
| stillbirth | 0.06538738425289344 |
| deaths | 0.05837143030522066 |
| heatstroke | 0.0723406981810143 |
| hypothermia | 0.06759553452278093 |
| lethal | 0.053420169875032834 |
| paramedic | 0.06759553452278093 |

## Socioeconomic impacts and resilience

### TF-IDF

|  |  |
| --- | --- |
| Word | TF-IDF Score |
| farmer | 0.051714250314898694 |
| occupational | 0.14836043419714431 |
| worker | 0.14592953822042465 |
| illness | 0.05006502103659244 |
| income | 0.10151705426658907 |
| deprive | 0.12795729600840808 |
| livelihood | 0.06470569585095608 |
| disease | 0.0997973954679797 |
| healthcare | 0.07660774881365272 |
| sanitation | 0.06210212887284381 |
| livestock | 0.16022573109431032 |
| elderly | 0.09758879878890356 |
| injury | 0.10747319405196093 |
| morbidity | 0.10847182909406783 |
| workplace | 0.2169930468564317 |
| outdoor | 0.08761378681403677 |
| fields | 0.055365328137724226 |
| alzheimer | 0.06752722968524305 |
| dehydration | 0.060430066323165034 |
| woman | 0.18958681148616016 |
| young | 0.11229947791056644 |
| homeless | 0.11270309564459428 |
| infarction | 0.05568126112419604 |
| glycaemic | 0.07937319824629005 |
| gastroenteritis | 0.060597782292308784 |
| neurologic | 0.06752722968524305 |
| parkinson | 0.05568126112419604 |
| physiopathological | 0.07937319824629005 |
| pharmacological | 0.06752722968524305 |
| psychiatric | 0.08767058512629805 |

## Disaster risk management

### TF-IDF

|  |  |
| --- | --- |
| Word | TF-IDF Score |
| building | 0.05670196378477415 |
| technology | 0.05517666726856947 |
| tree | 0.06876960174742022 |
| fire | 0.055894519734329205 |
| warning | 0.0640342369173302 |
| governance | 0.24380982709007062 |
| policy | 0.05892038599336493 |
| coal | 0.07413123874727293 |
| sensor | 0.07316855042618955 |
| overheating | 0.05211152435876629 |
| wildfire | 0.06414828310604605 |

## Risk identification and assessment

### TF-IDF

|  |  |
| --- | --- |
| Word | TF-IDF Score |
| cooling | 0.10687234458119961 |
| infrastructure | 0.1757889136604962 |
| simulation | 0.10763974106035648 |
| rainwater | 0.05466904957231441 |

## Urban risk and planning

### TF-IDF

|  |  |
| --- | --- |
| Word | TF-IDF Score |
| canopy | 0.08297571522058646 |
| roof | 0.0768514715681424 |
| overheat | 0.12336100801628415 |
| settlement | 0.12228382698373619 |
| property | 0.3842293633190632 |
| cool | 0.051353785807703764 |
| indoor | 0.06822806202343508 |
| vulnerability | 0.05099943697521045 |
| policymaking | 0.06587985160273914 |
| crime | 0.07726147185663887 |
| financing | 0.17756418058125203 |

# **References**

Bayomi, N., & Fernandez, J. E. (2023). Quantification of heat vulnerability using system dynamics. *Frontiers in Built Environment*, *9*, 1025480. https://doi.org/10.3389/fbuil.2023.1025480

ENBEL. (2023). (policy brief). *Protecting the vulnerable from from heat waves: a gap in policy* (No. 5).

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

García-León, D., Casanueva, A., Standardi, G., Burgstall, A., Flouris, A. D., & Nybo, L. (2021). Current and projected regional economic impacts of heatwaves in Europe. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-26050-z

Grootendorst, M. P. (2024). *Bertopic*. BERTopic - BERTopic. https://maartengr.github.io/BERTopic/api/bertopic.html#bertopic.\_bertopic.BERTopic.visualize\_topics\_per\_class

IPCC, Geneva, Switzerland, pp. 1-34, doi: 10.59327/IPCC/AR6-9789291691647.001

Jin, S., Stokes, G., & Hamilton, C. (2023). Empirical evidence of urban climate adaptation alignment with sustainable development: Application of LDA. *Cities*, *136*, 104254. https://doi.org/10.1016/j.cities.2023.104254

Kiarsi, M., Amiresmaili, M., Mahmoodi, M. R., Farahmandnia, H., Nakhaee, N., Zareiyan, A., & Aghababaeian, H. (2023). Heat waves and adaptation: A global systematic review. *Journal of Thermal Biology*, *116*, 103588. https://doi.org/10.1016/j.jtherbio.2023.103588

Koppe, C., Kovats, S., Jendritzky, G., & Menne, B. (2004). (rep.). *Heat-waves: risks and responses* (Ser. 2). WHO.

*Latest additions*. Latest additions | Knowledge base. (n.d.). Prevention Web. https://www.preventionweb.net/knowledge-base/type-content/latest-additions

Li, F., Yigitcanlar, T., Nepal, M., Thanh, K., & Dur, F. (2022). Understanding Urban Heat Vulnerability Assessment Methods: A PRISMA Review. *Energies*, *15*(19), 6998. https://doi.org/10.3390/en15196998

Mallick, T., Murphy, J., Bergerson, J. D., Verner, D. R., Hutchison, J. K., & Levy, L.-A. (2024). *Analyzing Regional Impacts of Climate Change using Natural Language Processing Techniques* (arXiv:2401.06817). arXiv. http://arxiv.org/abs/2401.06817

Meade, R. D., Notley, S. R., Akerman, A. P., McCormick, J. J., King, K. E., Sigal, R. J., & Kenny, G. P. (2023). Efficacy of Cooling Centers for Mitigating Physiological Strain in Older Adults during Daylong Heat Exposure: A Laboratory-Based Heat Wave Simulation. *Environmental Health Perspectives*, *131*(6), 067003. https://doi.org/10.1289/EHP11651

Muhamad Gatot Supiadin & Arif Dwi Laksito. (2023). Evaluating LDA and LSA for Topic Modeling in the Indonesian Natural Disaster. *Indonesian Journal of Computer Science*, *12*(6). https://doi.org/10.33022/ijcs.v12i6.3478

Ridgway, R. (2023). Choice and control: Corpus-based discourse analysis of teacher education policy in England (2010-2021). *Cogent Education*, *10*(1), 2212118. https://doi.org/10.1080/2331186X.2023.2212118

Safaei, M., & Longo, J. (2024). The End of the Policy Analyst? Testing the Capability of Artificial Intelligence to Generate Plausible, Persuasive, and Useful Policy Analysis. *Digital Government: Research and Practice*, *5*(1), 1–35. https://doi.org/10.1145/3604570

Sodoge, J., Kuhlicke, C., & De Brito, M. M. (2023). Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. *Weather and Climate Extremes*, *41*, 100574. https://doi.org/10.1016/j.wace.2023.100574

Stone, B., Gronlund, C. J., Mallen, E., Hondula, D., O’Neill, M. S., Rajput, M., Grijalva, S., Lanza, K., Harlan, S., Larsen, L., Augenbroe, G., Krayenhoff, E. S., Broadbent, A., & Georgescu, M. (2023). How Blackouts during Heat Waves Amplify Mortality and Morbidity Risk. *Environmental Science & Technology*, *57*(22), 8245–8255. https://doi.org/10.1021/acs.est.2c09588

Srivastava, B., Tafere, K., & Behrer, P. (2024). (working paper). *High temperature and learning outcomes evidence from Ethiopia*. The World Bank.

Tan, H., Kotamarthi, R., Wang, J., Qian, Y., & Chakraborty, T. C. (2023). Impact of different roofing mitigation strategies on near-surface temperature and energy consumption over the Chicago metropolitan area during a heatwave event. *Science of The Total Environment*, *860*, 160508. https://doi.org/10.1016/j.scitotenv.2022.160508

*Themes*. Themes and issues. (n.d.). Prevention Web.https://www.preventionweb.net/knowledge-base/themes#311

Văduva, A.-G., Munteanu, M., Oprea, S.-V., Bâra, A., & Niculae, A.M. (2023). Understanding Climate Change and Air Quality Over the Last Decade: Evidence From News and Weather Data Processing. *IEEE Access*, *11*, 144631–144648. https://doi.org/10.1109/ACCESS.2023.3345466

Volkanovska, E., Tan, S., Duan, C., Bartsch, S., & Stille, W. (2023). The InsightsNet Climate Change Corpus (ICCC): Compiling a Multimodal Corpus of Discourses in a Multi-Disciplinary Domain. *Datenbank-Spektrum*, *23*(3), 177–188. https://doi.org/10.1007/s13222-023-00454-1

Zander, K. K., Rieskamp, J., Mirbabaie, M., Alazab, M., & Nguyen, D. (2023). Responses to heat waves: What can Twitter data tell us? *Natural Hazards*, *116*(3), 3547–3564. https://doi.org/10.1007/s11069-023-05824-2

Zengul, F., Bulut, A., Oner, N., Ahmed, A., Yadav, M., Gray, H. G., & Ozaydin, B. (2023). *A Practical and Empirical Comparison of Three Topic Modeling Methods Using a COVID-19 Corpus: LSA, LDA, and Top2Vec*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2023.116

Zha, W., Ye, Q., Li, J., & Ozbay, K. (2023). A social media Data-Driven analysis for transport policy response to the COVID-19 pandemic outbreak in Wuhan, China. *Transportation Research Part A: Policy and Practice*, *172*, 103669. https://doi.org/10.1016/j.tra.2023.103669

Zhu, Y., Wang, J., Yuan, Y., Meng, B., Luo, M., Shi, C., & Ji, H. (2024). Spatial heterogeneities of residents’ sentiments and their associations with urban functional areas during heat waves– a case study in Beijing. *Computational Urban Science*, *4*(1), 7. https://doi.org/10.1007/s43762-024-00119-z

1. Some of the documents contain foreign words and characters mainly because of the location & author names as well as special terminology used. [↑](#footnote-ref-1)
2. Each document’s summary was examined by the researcher to check for appropriateness for inclusion to the corpus, even though they all exist under the same hazard category. [↑](#footnote-ref-2)
3. Involves documents related to identifying public health risks and ways to prevent them. [↑](#footnote-ref-3)
4. Involves documents related to the socioeconomic (e.g. economy, education, etc.) impacts, vulnerable groups, and ways to increase resilience. [↑](#footnote-ref-4)
5. Involves documents related to mainly infrastructural impacts in urban areas and ways to alleviate them. [↑](#footnote-ref-5)
6. Involves documents that identify & measure multiple risks. [↑](#footnote-ref-6)
7. Involves documents that identify risks and propose solutions to mitigate and adapt. May overlap with the Risk identification and assessment category. [↑](#footnote-ref-7)
8. Defines the minimum number of documents required to consider a topic meaningful and interpretable. [↑](#footnote-ref-8)
9. Defines the number of top words to extract for each topic. [↑](#footnote-ref-9)
10. Defines the range of n-grams to consider when generating features from the textual data. This is important for including words with multiple components. [↑](#footnote-ref-10)
11. Defines the probabilities for each topic, which is always set to “True”. [↑](#footnote-ref-11)
12. Please refer to Annex I for details regarding the workflow. Please refer to the workflow before examining the results. [↑](#footnote-ref-12)
13. The occurrences are filtered from the TF-IDF scores that are above the threshold and overlap with domain knowledge. [↑](#footnote-ref-13)
14. Please refer to Annex II. [↑](#footnote-ref-14)
15. The first topic is ‘Topic -1’. [↑](#footnote-ref-15)
16. The coherence scores and word probabilities are rounded for every category. [↑](#footnote-ref-16)
17. The reason for selecting 0.5 as the threshold is because this is the common practice. [↑](#footnote-ref-17)
18. The word lists are truncated. [↑](#footnote-ref-18)
19. The word lists are truncated. [↑](#footnote-ref-19)
20. The word lists are truncated. [↑](#footnote-ref-20)
21. [↑](#footnote-ref-21)
22. The tables do not depict all the TF-IDF score above the treshold. [↑](#footnote-ref-22)