“Un” Resilience: Drawing Insights from societal collapse

# introduction

Despite the extensive literature on resilience, the field often relies on generalized descriptions (Kirmayer et al., 2011), making it challenging to operationalize and measure resilience (Copeland et al., 2020). This difficulty may arise from the inherent challenge of directly observing resilience. When a system efficiently absorbs specific shocks, its observable resilience may be obscured by its absorptive capacity. For instance, if someone contracts COVID-19 but remains asymptomatic, their resilience to the virus remains unobserved and unmeasured.

Resilience discourse, widely applied across domains (Luthar, 2015), includes societal resilience defined by communities’ rebound capacity from crises (Fleming & Ledogar, 2008), encompassing resistance, recovery, and creativity (Maguire & Hagan, 2007). Scholars delineate social resilience into coping, adaptive, and transformative capacities. (Keck & Sakdapolrak, 2013). Essential for navigating challenges like climate change and urban expansion, quality of life indicators are crucial for a society’s ability to navigate uncertainty and adapt to shocks (Broch, 2013). However, resilience assumptions include linearity, distinct “stages” of absorption and recover, and “direct” system properties, with consensus on resilience definitions. While the core idea of resilience remains—bouncing back from stress or shock (Fleming & Ledogar, 2008) –varied perspectives challenge the identification of new contributions, requiring reconciliation.

Amid debates on identifying new contributions, a shift toward studying collapse emerges as an alternative perspective to understand societal responses to shocks. The ability to observe “collapse” opens new doors to look at a diversity of shock types—including financial, ecological, natural disasters, human displacement, and other socio-ecological disasters—of varying degrees and intersecting with one another to measure the collapse of society (even low-levels of “collapse,” meaning certain subsystems but not others).

Peterson posits, “Collapse is the shadow of resilience; consequently, studying collapse is indirectly the study of what makes a system resilient” (*Understanding Collapse, in Stockholm Resilience Centre - Research.*, 2018). In this project, we propose to study the collapse of a society, which is observable through criteria established by Cumming and Peterson (2017), as a way to infer the typology of shocks that lead to collapse (observable) or resilience (potentially unobservable). While it might also be difficult to observe societal “collapse,” we rely on a proxy for identifying candidate global events by using perceived societal crises through news coverage.

For decades, scholars have focused on studying socio-ecological collapse, emphasizing that “collapse” does not imply dark, apocalyptic events. Debates exist on “how much and what kind of change constitutes collapse,” but criteria have been developed to systematically observe it (Cumming & Peterson, 2017). Cumming and Peterson explain, “Collapse and resilience are two sides of the same coin; collapse occurs when resilience is lost, and resilient systems are less likely to collapse” (Cumming & Peterson, 2017). The four criteria to identify societal collapse are socio-ecological systems losing service or identity, rapid loss of system identity, significant capital loss, and enduring impacts (Cumming & Peterson, 2017).

Based on the above, we will systematically measure the public perception of societal collapse using existing data sources. Employ Natural Language Processing (NLP) and Machine Learning (ML) techniques to extract insights from news articles contextualized by data collected in surveys and interviews. This multifaceted approach aims to characterize the interconnections of shock-types. The data will help to explain perceptions of and propensity for societal collapse. We purposefully avoid studying the system’s specific configuration during the collapse, focusing instead on high-level failures (e.g., economic system broadly, food systems broadly—not each subsystem's configuration) caused by shocks/stressors. Considering specific factors wouldn’t significantly enhance the state of the art using novel methodologies toward understanding resilience.

A *crisis* is an event that could potentially lead to significant harm, and there are three necessary conditions that, taken together, are sufficient for its occurrence: (1) suddenness (sharp nonlinearity), (2) impact on a large population, and (3) significant harm to the well-being of that population within a relatively short period of time (Homer-Dixon et al., 2015). This definition assumes that the harm is a direct result of the crisis and, therefore, occurs relatively rapidly. It thus does not exclude the possibility that a crisis can precipitate widely beneficial changes in psychological states, social structures, and general well-being over a longer period of time.

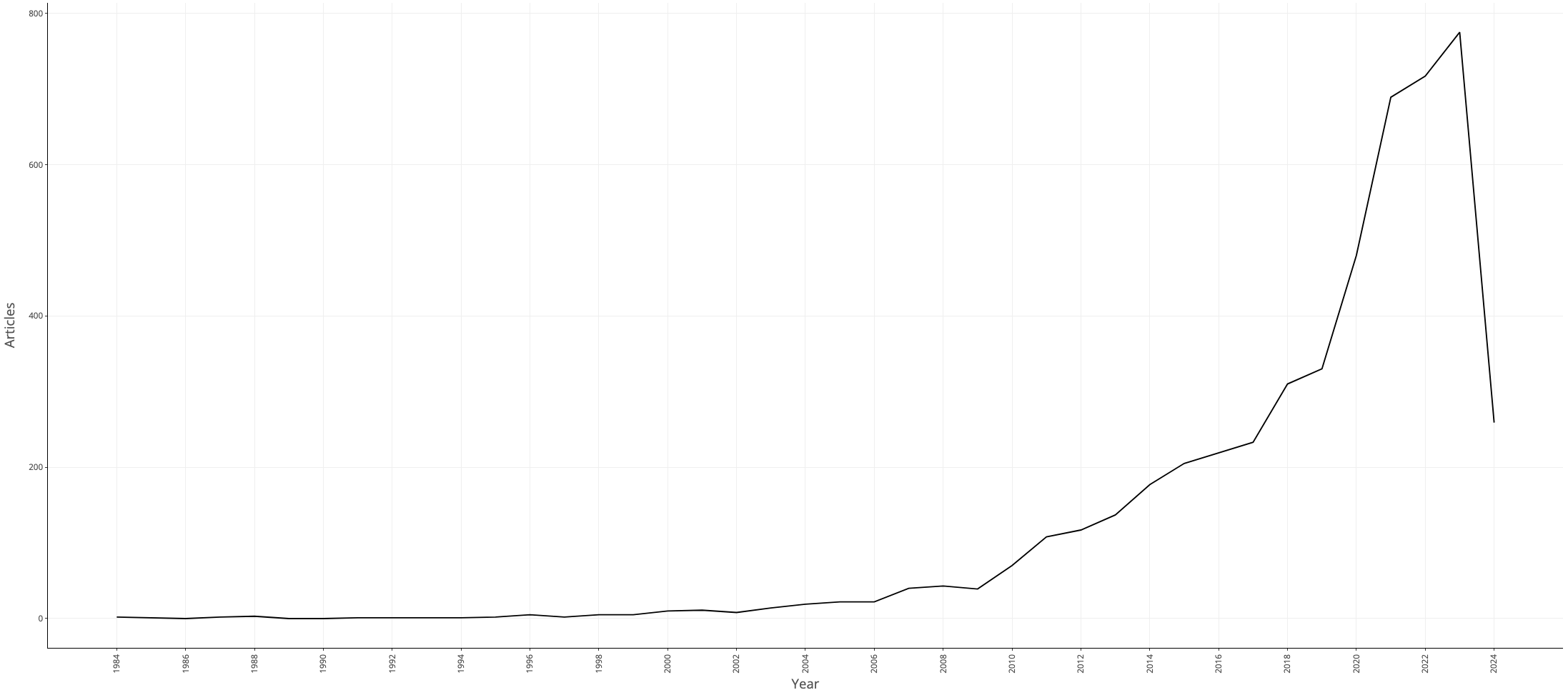
On the other hand, a *disaster* is defined as a significant disturbance to the operation of a community or society, resulting in extensive human, material, economic, or environmental losses and repercussions (UNDRR, 2009).The term "disaster" is thus only applied when there are long term and significant effects on social structures, including loss of life among human beings and destruction of property and related assets (Aubrecht et al., 2013). For example, the idea of a "natural" disaster suggests that it is triggered by a hazardous event of "natural" origin. However, the effects that ultimately qualify the label of "disaster" are largely the result of socio-economic conditions (Cannon, 2008; Quarantelli, 1995).

# Literature Review

"social resilience" AND ( shock OR crisis OR event ) AND NOT ( "mental health" OR pharmaceutical OR aviation )

Using the query “social AND resilience AND ( system OR community ) AND ( shock OR crisis OR event )”, we could extract 5,136 documents in Scopus. The subject areas of the documents are: Social Sciences (2,037); Environmental Science (1,433); Medicine (1,113); Engineering (836); Earth and Planetary Sciences (600); Computer Science (440); Psychology (425); Business, Management and Accounting (403); Agricultural and Biological Sciences (395); Energy (308); Economics, Econometrics and Finance (280); Arts and Humanities (276); Nursing (145); Mathematics (142); Decision Sciences (130); Neuroscience (108); Biochemistry, Genetics and Molecular Biology (107); Multidisciplinary (96); Health Professions (69); Materials Science (43); Physics and Astronomy (39); Immunology and Microbiology (28); Pharmacology, Toxicology and Pharmaceutics (26); Chemical Engineering (25), Chemistry (16); Veterinary (15).

We will proceed to do a bibliometric analysis using bibliometrix (Aria and Cuccurullo, 2017) and could analyze 5,085 documents. The documents cover the search from 1984 to 2024. In the Figure, we can observe how those keywords gain relevance through the years, reaching the highest in 2023.



# Definition of concepts

The *collapse* has been defined from different social, economic, and social perspectives, specifically, as a rapid loss of an established social, political, and economic complexity (Tainter, 1988). Hanson and Rees (2008) defined collapse from a social perspective in which productivity is reduced, impacting social systems rapidly and causing significant disruptions that provoke social change. Abel et al. (2006) developed a comprehensive concept of collapse that incorporates different social and ecological perspectives and allows to address resilience and vulnerability, as in the case of socio-ecological collapse, understood as “(...) major losses of social, human, and natural capitals through the breakdown of social networks, deaths of individuals, loss of knowledge, depletion of flora and fauna for food and medicine, and loss of access to ceremonial sites and lands”.

When addressing the concept of collapse, other terms such as social-ecological system, event, shock, disaster, and crisis are inherently related and are defined in this section. A *system* is an interconnected set of elements that is coherently organized in a way that achieves a purpose (Ghosh, 2015). The elements making up human social-ecological systems include resources, nonhuman organisms, people, organizations, institutions, and technologies (Beddoe et al., 2009).

*Events* are defined as situations that occur within a certain timeframe and involve entities or objects (Rodrigues & Abel, 2019). Complex systems, characterized by irregular and non-periodic dynamics, often give rise to extreme events with significant impact (Kantz, 2010). These events are the result of the interactions between internal functional subsystems, which operate at their own rhythms and can lead to the emergence of higher-order complex events (Ehresmann & Vanbremeersch, 2011).

In the context of political events, the World Handbook of Political Indicators IV (WHIV)[[1]](#footnote-2) provides a set of country-level measures for contentious political events and identifies particular components in an event, such as an actor, an event form or type, and one or more targets, along with the date and location (Jenkins et al., 2012). WHIV contains 40 event forms but could be resumed into four main categories of event types such as protest, political violence, political sanctions and political relaxations

*Shock* can be defined as a sudden, significant disruption that challenges the system's ability to maintain its functionality and adapt to new circumstances (Srinivasan & Kumar, 2015). These shocks can be caused by a variety of factors, including natural disasters, human-made events, and changes in environmental conditions (Jenelius & Mattsson, 2020). Pradhan and Mukherjee (2018) cite Clarke and Dercon (2015), who categorized shocks into various types depending on their origins, such as climate-related, economic, political, criminal, and health-related.

Authors like the Organization for Economic Co-operation and Development (OECD) (2014) and Pradhan and Mukherjee (2018) mention that at a macro level, shocks can be categorized into *covariate shock* and *idiosyncratic shock*. Covariate shock refers to events that affect entire nations or communities, encompassing natural disasters and economic fluctuations like price increases and economic downturns (Feeny & Miller-Dawkins, 2016). Idiosyncratic shock refers to events that specifically affect individuals and families, such as the death of the main breadwinner or the loss of income-generating activity (Feeny & Miller-Dawkins, 2016).

A *crisis* is an event that could potentially lead to significant harm, and there are three necessary conditions that, taken together, are sufficient for its occurrence: (1) suddenness (sharp nonlinearity), (2) impact on a large population, and (3) significant harm to the well-being of that population within a relatively short period of time (Homer-Dixon et al., 2015). This definition assumes that the harm is a direct result of the crisis and, therefore, occurs relatively rapidly. It thus does not exclude the possibility that a crisis can precipitate widely beneficial changes in psychological states, social structures, and general well-being over a longer period of time.

On the other hand, a *disaster* is defined as a significant disturbance to the operation of a community or society, resulting in extensive human, material, economic, or environmental losses and repercussions (UNDRR, 2009).The term "disaster" is thus only applied when there are long term and significant effects on social structures, including loss of life among human beings and destruction of property and related assets (Aubrecht et al., 2013). For example, the idea of a "natural" disaster suggests that it is triggered by a hazardous event of "natural" origin. However, the effects that ultimately qualify the label of "disaster" are largely the result of socio-economic conditions (Cannon, 2008; Quarantelli, 1995).

# characterizations of events

An event is an action that has an impact on the system. These actions are usually performed by an actor over another actor, but in some scenarios, they are performed by no actors. Since an event can cause or influence another event, there is also the possibility that several events can have causal connections.

Several events can belong to the same category, but each individual event has a different impact on the system. The magnitude of that impact is the differentiator between two different systems. For instance, an increase of 1% in inflation in Venezuela might not impact the country as much as that same 1% might impact the United States. However, how to measure that impact is tricky, specially when dealing of societal impact. One way of addressing it is identifying how many subsystems were affected.

# Proposed Method

The research relies on using “inference from absence” or “negative inference” as an indirect/inverse way to measure the resilience of systems (Wallach, 2019). This approach has been used in various fields, including anthropology/paleontology, environmental sciences, and physics, where direct observation and/or measurement of a phenomenon is difficult or impossible. By focusing on societal collapse/crisis and characterizing interconnected shocks that led to collapse, we can provide insights about those societies experiencing shocks that did not collapse or fall into crisis.

The research methodology is divided into four phases. In Phase 1, we focus on identifying potential "candidate" shock events. In Phase 2, the emphasis shifts toward identifying those candidate shocks that can cause the system to collapse. Phase 3 involves analyzing the resilience of the collapsed systems and how they recovered. Finally, in Phase 4, the focus is on characterizing the attributes of the shocks that lead to collapse or crisis.

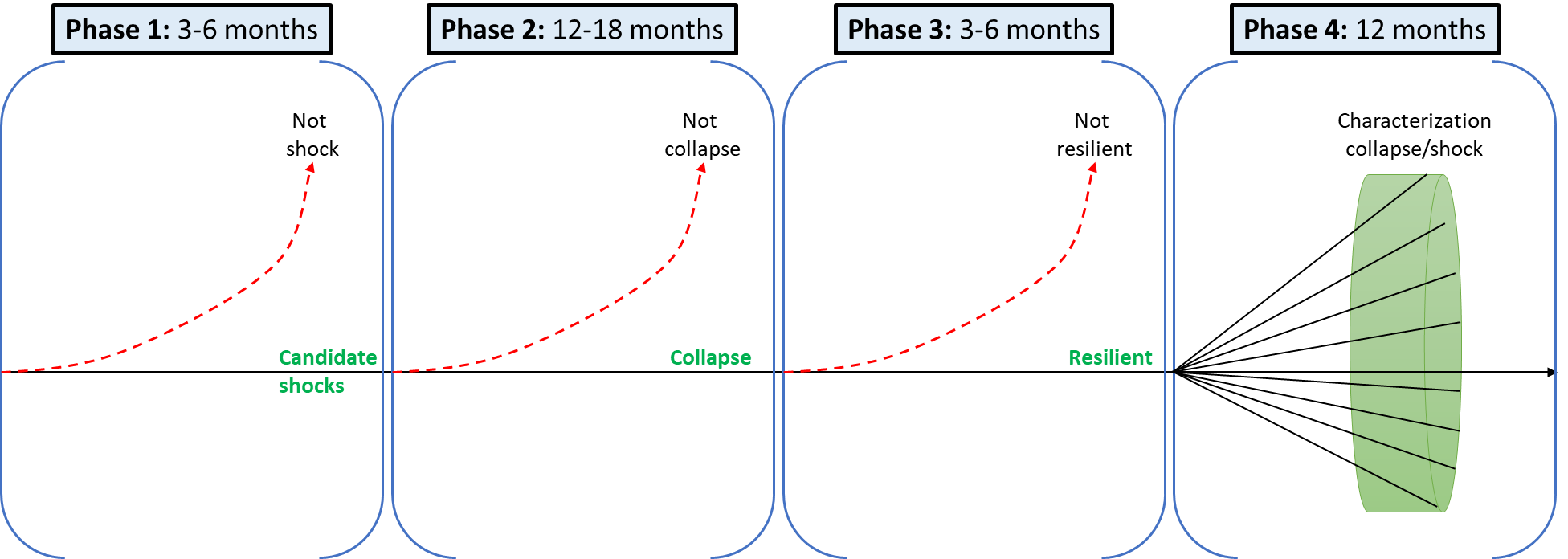
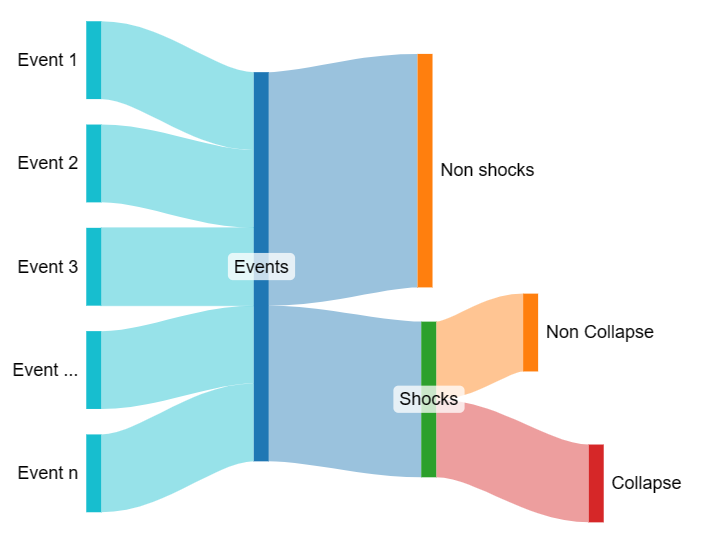


Figure 1. Research Proposal.



**Resilience 1:** The system NOT collapsing.

**Resilience 2:** The collapsed system coming back to non-collapse.



**Research Questions:** What shocks, including subsequent or co-occurring shocks, lead to social perceptions of crisis/disaster? How do we systematically identify societal collapse? What characterization of shock types and combinations leads to perceptions of societal collapse? How does the typology of shock types and combinations explain the propensity for societal collapse? What are the perceived mechanisms that fail?

# Phase 1: Identify cases of societal collapse/ “un” resilience

Adapting existing crisis/disaster lexicons used to study social and news media, we will employ NLP and ML algorithms to identify "candidate “events” related to societal collapse in global news databases. These identified candidate events will be stored and named as a "Global Candidate Shock News Database" (GCSND) and will serve as a centralized source of information on events to be analyzed in phase 2. This frames our study as the social perception of crisis/collapse as conveyed through media. NLP/ML will also assist in identifying “shocks” that preceded the crisis/collapse event. Human coders will be used to verify that the NLP/ML correctly identifies events and shocks. In this dataset, we will also track aspects of the shocks, including the length of time the shock remained in public conversations, volume/number of national or international news outlets covering the shock, tone and sentiment of coverage, and mentions of actors, institutions, systems, or services that ceased functioning.

## “Global Candidate Shock News Database” (GCSND) (Shock or not shock)

The framework contains five steps to build the GCSND, as shown in Figure 2. Step 1 is to identify the lexicons associated with crisis/shock. A base lexicon is going to be a point start. Then, it is important to identify the lexicons that help search for a specific type of shock. If the sets of lexicons do not capture the type of shocks that we are looking for, the base lexicons are used as the lexicons found to search news articles in the data sources. Step 2 is to extract metadata and content from news articles. Step 3 is to identify the news's topics or “candidate” events and then create a database with the data extracted (shocks, metadata, content, events).

A diagram of different types of objects

Description automatically generated

Figure 2. Framework for building a database of news articles.

### Step 1: Collapse Lexicon

A lexicon-based approach has been used to analyze financial (Schröter & Storjohann, 2015), political (Havumetsä, 2023; Zouave, 2014), and social (Jayanetti et al., 2023; Salem et al., 2022) crises. Appropriately using lexicons for searching news is crucial because they extract valuable information from unstructured data sources like online news stories (Shaikh et al., 2019).We have identified a lexicon created for social media related to crises that we will use for search in news articles: CrisisLex, a lexicon of 380 automatically generated and human-curated terms frequently related to disasters (Olteanu et al., 2014a). However, they did not define what a crisis is and were limited to capturing only natural disasters and human-made disasters, excluding other types of crises/events such as financial, ecological, natural disasters, human displacement, and other socio-ecological disasters.

Given those limitations, more lexicons must be included for other types of crises that might cause collapse. We propose to use the GDELT database to examine different events reported in the media and extract lexicons related to those crises/events. Jayanetti et al. (2023) used GDELT to collect data on unique events, where they identified eight main GKG themes (Project, 2021) related to xenophobic events toward the refugee population. Additionally, more keywords can be added using different approaches to extract relevant keywords (Habibi & Popescu-Belis, 2015; Olteanu et al., 2014b) to make the lexicon more robust. All of these will be combined into a collapsed lexicon.

We will select significant incidents such as COVID-19, the Russian-Ukraine conflict, Hurricane Maria, and the 2007-2008 financial crisis to determine whether the lexicons can capture various events and crises. We aim to verify whether news related to these events can be extracted using the collapsed lexicon generation. This lexicon will be used to extract newspapers that contain its words using binary search. The keywords of this lexicon will be clustered using affinity propagation (Dueck, 2009) to identify groups of similar words and, using existing LLMs, assign the factor names. These groups can represent a measurable variable, where its value can be the sum of the frequency of all the words in the group. Human coders will be used to verify that the NLP/ML correctly identifies the appropriate group (considering that a keyword can be associated with multiple types of events) and the appropriate name of the factor names. CollapseLexicon CollpaseLexicon

### Step 2: Searching “Candidate” News/Events

We will use the Lexicon from Step 1 as a search query to extract the content and metadata of the news articles from local, regional, national, and international outlets with Quantexa News API[[2]](#footnote-3). This API allows us to search news outlets in all countries from 1924 (claimed) to the present using keyworks and parameters like country, language and sentiment; likewise, to extract the text and other parameters like date, title, sentiment, language, country, etc. While the pricing needs to be discussed with a sales representative, discussions online mention a starting plan of 49 USD per 30,000 articles.

Metadata is structured information (variables) of a digital object and it is as essential as the content itself (Khan et al., 2023). With the content we will identify significant events, key individuals, and their relationships with other entities; and with the metadata, we will create a Shock News Metadata (SNM) database of the news. Some other important ones for extracting from news websites are highlighted in a red square in Figure 3. For instance, the variable publication date will allow us to establish the duration of media coverage, the number of national or international news outlets covering the event over time, the tone, and the sentiment of the content of the news.

**Content Analysis**

We will identify events, shocks, …, [simple narrative]. And then per event, do the analyses we mentioned.

To identify “candidate” events, we will analyze the text from the news articles, which will be qualitatively coded using content analysis. Content analysis is a method used in various disciplines to extract quantitative measures from textual information. It involves aggregating, processing, and analyzing qualitative data to derive meaningful insights (Donald, 2022; Mir et al., 2018; Oleinik, 2021; Wang et al., 2022). This allows us to determine the presence of key entities (actors, institutions, systems, services, among others) that ceased functioning, factors, mechanisms, certain tokens, topics/events, and the relationship between a set of news that generates an event within a given qualitative data.

Based on Lydia’s architecture[[3]](#footnote-4) (Lloyd et al., 2005), which is a system that can track the temporal and spatial distribution of the entities in the news: who is being talked about, by whom, when, and where? We can analyze different entities from the content of the news using these four steps: 1) identify key entities, 2) for each entity, identify what other entities occur near it, 3) set a synonym identification (i.e., George Bush can be referred to as Bush, President Bush, and George W. Bush), and 4) temporal and spatial analysis.

We will implement a named entity recognition model (NER) to identify key entities. In a previous study, we combined interviews and web scraping to gather data on key actors and their relationships in the humanitarian response. We employed Social Network Analysis (SNA) to examine these relationships and the collaborative strategies among entities, focusing on how they address the migration challenges (Romero et al., 2024). We can generalize this study to macro/global events and analyze the relationships between key actors during different events.

Fig 5 is an example on how LLM embedding can be used to identify what a set of news are talking about

A screen shot of a computer screen

Description automatically generated

Figure 5. Topic modeling of social crisis using Nomic Atlas: News related to Venezuelan migration in Colombia.

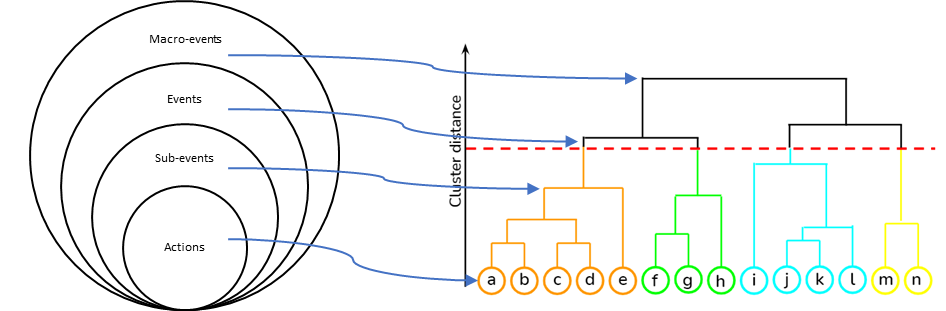
[Diagram on how we can use social media for content analysis]

### Step 3: Events and sub-events, and macro-events identification

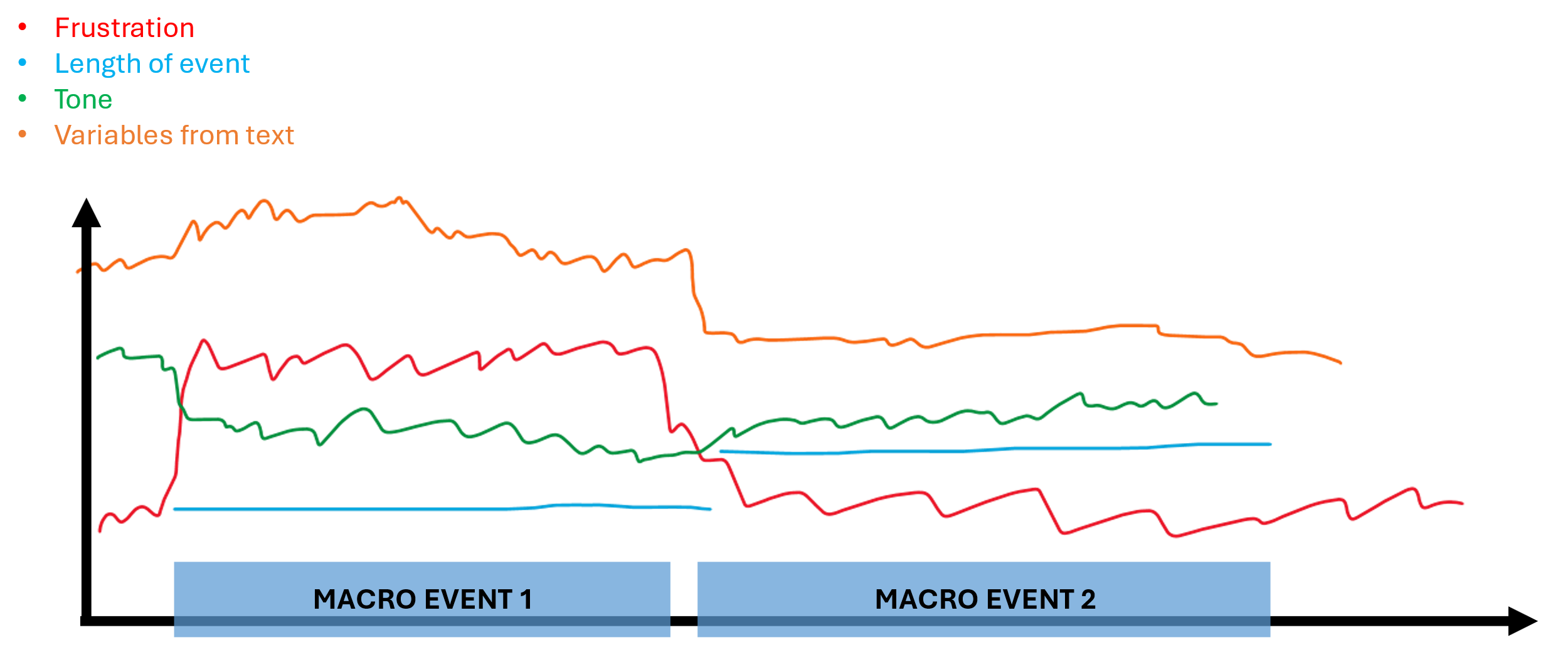
There is a hierarchy of events, considering that some can be aggregated into a more general event. Actions are the smallest level of aggregation, based on GDELT, is anything that happens between two actors.

Actions would be grouped into sub-events and these into events using clustering:

* Affinity propagation (Dueck, 2009) (used in the tests with clustering the lexicon)
* SHARI is a method that combines various tools with web archive collections to analyze and visualize the most significant news story for a specific date (Jones et al., 2020). [StoryGraph](https://storygraph.cs.odu.edu/) groups news articles to pinpoint a prevailing narrative. Hypercane[[4]](#footnote-5) utilizes ArchiveNow to archive URLs generated by StoryGraph (Handcock & Jones, 2004). Hypercane then analyzes these URLs to identify prevalent terms, entities, and high-quality images for social media storytelling (Handcock & Jones, 2004). [Raintale](https://oduwsdl.github.io/raintale/) synthesizes the results from these tools to visually represent the news story for a particular day. This integrated process is called SHARI (StoryGraph Hypercane ArchiveNow Raintale Integration) (Jones et al., 2020; Jones et al., 2021)
* Atlas[[5]](#footnote-6)
* Sub-event detection (Pohl et al., 2012)



### Step 4: Alignment into a timeline of events + variables + indicators +…



These elements will be organized through time:

* the list of events from phase 3
* the values of the variables from phase 1 and others like frustration, tone, …
* macro-level indicators such as GDP, poverty index, education index, …

### Step 5: Identification of “candidate” events → Shock/not shock



Criteria for an event for being considered as candidate. Like a threshold or a shock score

A Change point analysis will be run for the variables that were identified in the phase 1 to identify when there are abrupt changes. And those events around that change point time will be joined as candidate events

# Phase 2: Statistical analysis of societal indicator data

The comparison between cases using indicators data from the World Bank and Statista (longitudinal) and survey data (Phase 4, cross-sectional) will allow for an understanding of the conditions before, during, and after the crisis/collapse in each context. The responses to the same event (e.g., Covid-19) in different countries can vary in intensity based on those conditions. The perceived “collapse” can occur at different times after the event, depending on contextual factors. Statistical analysis of changepoint analysis using available public and subscription data will allow for an understanding of how these conditions interact for incorporation into Phase 3.

# Phase 3: Specifying the system to operationalize collapse/crisis

Based on the notion of system “identity” (Cumming & Peterson, 2017), which Cumming and Peterson argue is central to operationalizing the collapse of a system, we will use publicly available or subscription data to characterize the countries in which societal collapse/crisis occurred. Each of Cumming and Peterson’s “four criteria for defining collapse” will be mapped to available data. This includes:

* **Criterion 1:** loss of actors/institutions/services (unique events from Phase 1).
* **Criterion 2:** collapse speed (length of media coverage in Phase 1).
* **Criterion 3:** substantial loss of social-ecological capital (statistical analysis from Phase 2).
* **Criterion 4:** lasting consequences (ongoing coverage and qualitative data in Phase 4).

These four criteria will allow us to label the cases in the dataset as “collapse” or not. Human coders and subject matter experts will verify these.

### How to measure criterion 1 (identity of the system)

### How to measure criterion 2 (quick degradation)

### How to measure criterion 3 (loss of social-ecological capital)

### How to measure criterion 4 (lasting consequences)

# Phase 4: Characterize “Shocks” attributed to collapse/crisis

Conduct surveys and interviews in South America and Eastern Europe, led by our area-studies experts, to 1) validate the NLP/ML identification of shocks and collapse/crisis; 2) identify and explain the types of shocks related to local events selected from the dataset; 3) characterize the interconnectedness of identified shocks; and 4) identify/explain the societal long-term outcomes of those shocks. This phase relies on a mixed-method (content and statistical analysis) approach to developing 1) a typology of shocks, 2) a network structure model of shock interconnectedness/heterarchies [11], and 3) network measurements to characterize network structures of interconnected shocks.

# Phase 4: Develop a Theory of Societal Structure that Precipitates Collapse/Crisis.

Using the datasets generated in Phases 1-3, we will create a theory of societal collapse. The adverse inference [10] of societal collapse provides insights into societal resilience. In this case, we will connect the network typology of interconnected shocks (Phases 1 & 3) to the dataset of cases classified as “collapse/crisis” or not (Phases 1 & 2) to speak to commonalities and differences across those systems that collapsed and those that did not.

# complementary Data sources

Throughout each phase, we will leverage multiple data sources to extract the required information. Table 2 describes potencial data sources to support the proposal. Subsequently, we will detail their significance for our project and discuss how to use them.

Table 2 . Data sources description

|  |  |  |  |
| --- | --- | --- | --- |
| **Phase 1** | | | |
| **Type of source** | **Source** | **Pricing** | **Description** |
| Global News Databases | Google News | Free | News-gathering tool that selects and presents news stories from a wide range of sources (Schroeder & Kralemann, 2005). It presents a continuous flow of links to articles from thousands of publishers and magazines. |
| GDELT | Free | Global Database of Events, Location, and Tone (GDELT), is a real-time, large-scale database that monitors global news sources. Its coverage and real-time nature make it a valuable resource for understanding global human society and its interactions (Chen et al., 2016). |
| LexisNexis | Subscription | LexisNexis is a comprehensive online information retrieval system that provides access to a wide range of sources, including legal information, news, business magazines, and government documents (Scales & Gilles, 1995). |
| Newsapi.org |  | Limitation: Do not extract the content |
|  | Newsdata.io |  | Limitation: Do not extract news less than 2 years old. |
|  | Bing News Search |  | Limitation: Do not extract the content and it is too expensive. |
|  |  |  |  |
| Social Media | Twitter (X) | Subscription |  |
| Reddit | Free | h |
| Social Media Archive | Free |  |
| Meta Content Library | Free |  |
| Meta's Ad Targeting Transparency dataset | Free |  |
| Other sources | Web Archive | Free |  |
| Wikidata | Free |  |
| **Phase 2** | | | |
| **Type of source** | **Source** | **Pricing** | **Description** |
|  | World Bank | Free |  |
|  | Statista |  |  |

### Google News

Google News covers 141+ countries and 41+ languages. It allows automating the process of extracting information using the GNews Python API[[6]](#footnote-7). Additionally, the use of keywords let restricting results to a customizable time span. Each GNews query can return 100 results in JSON format.

The advantage of using Google News API is that we can extract different types of variables without doing a scrapping process. The following variables can be extracted through the API:

* Title: Title of the article
* Topic/Event: Context of the new
* URL: Link to the article (id of the new).
* Published date: Published date.
* Description: Short description of the article.



* Publisher: Publisher of the article.

Then, using the URL of the data, we can extract the content of the article:

* Title: Title of the article
* Text: Text of the full article
* Images: Images of the article
* Authors: Who wrote the article
* URL: Link to the article (id of the new)

### GDELT

The Global Data on Events, Location, and Tone (GDELT) Project (Leetaru & Schrodt, 2013) collects and processes news articles from around the world in near-real time. The service monitors hundreds of thousands of news outlets and live-translates articles from 65 languages to English (Project, 2015). In addition, the content of each article is processed and labeled to classify features such as events, themes, actors, locations, and sentiment. GDELT also provides a TV news dataset (Leetaru & Schrodt, 2013) that processes the closed captioning of broadcasts from 14 news stations into searchable n-grams. GDELT's TV news API (Project, 2017) provides an interface to the Internet Archive's TV News Archive, allowing TV news analysis from over 150 stations since 2009. These datasets are freely available through APIs and database tables hosted through Google's BigQuery service. Although the data is easily accessible, extracting data requires constructing potentially complex SQL queries. GDELT is a prime example of big data, often defined as requiring the "three Vs": variety, volume, and velocity (Gandomi & Haider, 2015). GDELT provides many features for each article, including measuring over 2300 emotions and themes through its Global Content Analysis Measures (GCAM) and event-coding based on the Conflict and Mediation Event Observations (CAMEO) framework. GDELT contains a large volume of data, with just the 2015 news article data requiring over 2.5 TB, and provides new data at a high velocity, with an update rate of every 15 minutes (Project, 2015).

Until 2013, does not include codes for financial events, disease outbreaks and natural disasters (Leetaru & Schrodt, 2013)

**Comments by Brian: Understanding about GDELT (Do not use in the proposal)**

GDELT used LexisNexis to extract news articles, focusing on three news wire agencies: 1) Agence France Presse, 2) Associated Press, and 3) Xinhua. Then, they analyzed the correlation between those agencies regarding time and geography. Excluded keywords like non-insignificant sports volume and financial coverage were retained. Such coverage can create complications for the event coding process in that it often contains language very similar to that used to describe violent political events, such as two sports teams battling it out or a company's stock price “under siege”.

To create the event database, we followed a series of steps. First, we downloaded all relevant content from each newswire using a specific query from LexisNexis. Then, we filtered out any sports- or finance-related news using a post-processing Boolean query. Next, we used a geocoding tool from Leetaru (2012) to identify and clarify any geographic references in the articles. After that, we applied the Tabari system to extract all events in each article and used the Tabari geocoding post-processing system to link each event to its specific location. The final list of events for each newswire is internally deduplicated. Multiple references to the same event across one or more articles from the same newswire are collapsed into a single event record. To allow the study of each newswire individually, events are not deduplicated across newswires (externally deduplicated).

### LexisNexis

Nexis[[7]](#footnote-8) database provides direct access to over 36,000 journals, regional, national, and international newspapers, company and business sources, and biographical information.

### Social Media Archive

Social Media Archive[[8]](#footnote-9):  Some datasets are **free**.

### Meta

Meta Content Library and API[[9]](#footnote-10). We need to request access, but if granted there is a lot of information that can be retrieved for **free**.

Meta's Ad Targeting Transparency dataset[[10]](#footnote-11): granular, ad-level targeting information for all *social issue*, *electoral*, and *political*ads run across Meta's technologies since August 2020 in 120+ countries. We need to apply, but it is **free**.

### Web Archive

The web archive is a valuable tool that facilitates the exploration of link sources that are absent on the web and, therefore, inaccessible for data mining. One of the best-known repositories for web archives is Wayback Machine[[11]](#footnote-12). It works as a digital archive of the World Wide Web and other resources on the Internet, allowing users to browse and access archived versions of web pages at different historical points in time.

In some cases, the news is not available on a public server; for this reason, if the content cannot be extracted, we can access the information using a web archive, specifically the Wayback Machine.

### Wikidata

Wikidata[[12]](#footnote-13) is a free and open knowledge base that humans and machines can read and edit. It acts as central storage for the structured data of its Wikimedia sister projects, including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.

REFERENCES

Abel, N., Cumming, D. H. M., & Anderies, J. M. (2006). Collapse and Reorganization in Social-Ecological Systems

Questions, Some Ideas, and Policy Implications. *Ecology and Society*, *11*(1). <http://www.jstor.org/stable/26267805>

Aubrecht, C., Fuchs, S., & Neuhold, C. (2013). Spatio-temporal aspects and dimensions in integrated disaster risk management. *Natural Hazards*, *68*(3), 1205-1216. <https://doi.org/10.1007/s11069-013-0619-9>

Beddoe, R., Costanza, R., Farley, J., Garza, E., Kent, J., Kubiszewski, I., Martinez, L., McCowen, T., Murphy, K., Myers, N., Ogden, Z., Stapleton, K., & Woodward, J. (2009). Overcoming systemic roadblocks to sustainability: The evolutionary redesign of worldviews, institutions, and technologies. *Proceedings of the National Academy of Sciences*, *106*(8), 2483-2489. <https://doi.org/10.1073/pnas.0812570106>

Broch, H. B. (2013). Social resilience-local responses to changes in social and natural environments. *Maritime Studies*, *12*(1), 6.

Cannon, T. (2008). Vulnerability,“innocent” disasters and the imperative of cultural understanding. *Disaster Prevention and Management: An International Journal*, *17*(3), 350-357.

Clarke, D., & Dercon, S. (2015). Insurance, credit and safety nets for the poor in a world of risk. *Financing for Overcoming Economic Insecurity*, 85-110.

Copeland, S., Comes, T., Bach, S., Nagenborg, M., Schulte, Y., & Doorn, N. (2020). Measuring social resilience: Trade-offs, challenges and opportunities for indicator models in transforming societies. *International journal of disaster risk reduction*, *51*, 101799.

Cumming, G. S., & Peterson, G. D. (2017). Unifying research on social–ecological resilience and collapse. *Trends in ecology & evolution*, *32*(9), 695-713.

Donald, W. E. (2022). Content analysis of metadata, titles, and abstracts (CAMTA): application of the method to business and management research. *Management Research Review*, *45*(1), 47-64.

Dueck, D. (2009). *Affinity propagation: clustering data by passing messages*. University of Toronto Toronto, ON, Canada.

Ehresmann, A. C., & Vanbremeersch, J.-P. (2011, 2011//). Analysis of Complex Events in Memory Evolutive Systems. Towards a Service-Based Internet. ServiceWave 2010 Workshops, Berlin, Heidelberg.

Feeny, S., & Miller-Dawkins, M. (2016). Household Vulnerability and Resilience to Economic Shocks in Melanesia: An Overview. *Household Vulnerability and Resilience to Economic Shocks*, 1-22.

Fleming, J., & Ledogar, R. J. (2008). Resilience, an evolving concept: A review of literature relevant to Aboriginal research. *Pimatisiwin*, *6*(2), 7.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

Ghosh, A. (2015). Thinking in Systems. In (pp. 1-17). Springer International Publishing. <https://doi.org/10.1007/978-3-319-10735-6_1>

Habibi, M., & Popescu-Belis, A. (2015). Keyword extraction and clustering for document recommendation in conversations. *IEEE/ACM Transactions on audio, speech, and language processing*, *23*(4), 746-759.

Handcock, M. S., & Jones, J. H. (2004). Likelihood-based inference for stochastic models of sexual network formation. *Theoretical population biology*, *65*(4), 413-422.

Hanson, R., & Rees, M. J. (2008). Catastrophe, social collapse, and human extinction. In N. Bostrom & M. M. Cirkovic (Eds.), *Global Catastrophic Risks* (pp. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780198570509.003.0023>

Havumetsä, N. (2023). Lexical borrowing in journalism in a time of political crisis. *Perspectives*, *31*(3), 562-575.

Homer-Dixon, T., Walker, B., Biggs, R., Crépin, A.-S., Folke, C., Lambin, E. F., Peterson, G. D., Rockström, J., Scheffer, M., Steffen, W., & Troell, M. (2015). Synchronous failure: the emerging causal architecture of global crisis. *Ecology and Society*, *20*(3). <http://www.jstor.org/stable/26270255>

Jayanetti, H. R., Frydenlund, E., & Weigle, M. C. (2023). Xenophobic Events vs. Refugee Population--Using GDELT to Identify Countries with Disproportionate Coverage. *arXiv preprint arXiv:2308.05038*.

Jenelius, E., & Mattsson, L.-G. (2020). Resilience of Transport Systems. In.

Jenkins, C. J., Taylor, C. L., Abbott, M., Maher, T. V., & Peterson, L. (2012). The World Handbook of political indicators IV: An introduction and codebook. *Mershon Center for International Security Studies, The Ohio State University*.

Jones, S. M., Nwala, A. C., Klein, M., Weigle, M. C., & Nelson, M. L. (2020). SHARI--An Integration of Tools to Visualize the Story of the Day. *arXiv preprint arXiv:2008.00139*.

Jones, S. M., Weigle, M. C., & Nelson, M. L. (2021). Hypercane: toolkit for summarizing large collections of archived webpages. *ACM SIGWEB Newsletter*, *2021*(Summer), 1-14.

Kantz, H. (2010). Dynamics and Statistics of Extreme Events. In (pp. 205-216). Springer London. <https://doi.org/10.1007/978-1-84996-396-1_10>

Keck, M., & Sakdapolrak, P. (2013). What is social resilience? Lessons learned and ways forward. *Erdkunde*, 5-19.

Khan, M., Alharbi, Y., Alferaidi, A., Alharbi, T. S., & Yadav, K. (2023). Metadata for Efficient Management of Digital News Articles in Multilingual News Archives. *SAGE Open*, *13*(4), 21582440231201368.

Kirmayer, L. J., Dandeneau, S., Marshall, E., Phillips, M. K., & Williamson, K. J. (2011). Rethinking resilience from indigenous perspectives. *The Canadian Journal of Psychiatry*, *56*(2), 84-91.

Leetaru, K., & Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. ISA annual convention,

Lloyd, L., Kechagias, D., & Skiena, S. (2005). Lydia: A system for large-scale news analysis. International Symposium on String Processing and Information Retrieval,

Luthar, S. S. (2015). Resilience in development: A synthesis of research across five decades. *Developmental psychopathology: Volume three: Risk, disorder, and adaptation*, 739-795.

Maguire, B., & Hagan, P. (2007). Disasters and communities: Understanding social resilience. *Australian Journal of Emergency Management, The*, *22*(2), 16-20.

Mir, S., Lu, S.-H., Cantor, D., & Hofer, C. (2018). Content analysis in SCM research: past uses and future research opportunities. *The International Journal of Logistics Management*, *29*(1), 152-190.

OECD. (2014). *Guidelines for Resilience Systems Analysis: How to Analyse Risk and Build a Roadmap to Resilience*. <https://doi.org/doi:https://doi.org/10.1787/3b1d3efe-en>

Oleinik, A. (2021). Uses of content analysis in economic sciences: An overview of the current situation and prospects. *VOPROSY ECONOMIKI*(4).

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014a). CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 376-385. <https://doi.org/10.1609/icwsm.v8i1.14538>

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014b). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. Proceedings of the international AAAI conference on web and social media,

Pohl, D., Bouchachia, A., & Hellwagner, H. (2012). Supporting crisis management via sub-event detection in social networks. 2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises,

Pradhan, K. C., & Mukherjee, S. (2018). Covariate and Idiosyncratic Shocks and Coping Strategies for Poor and Non-poor Rural Households in India. *Journal of Quantitative Economics*, *16*(1), 101-127. <https://doi.org/10.1007/s40953-017-0073-8>

Project, T. G. (2015). *GDELT 2.0: Our Global World in Realtime*. <https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>

Project, T. G. (2017). GDELT 2.0 Television API Debuts! <https://blog.gdeltproject.org/gdelt-2-0-television-api-debuts/>

Project, T. G. (2021). *New November 2021 GKG 2.0 Themes Lookup.* <https://blog.gdeltproject.org/new-november-2021-gkg-2-0-themes-lookup/>

Quarantelli, E. L. (1995). Editor's Introduction: What Is a Disaster? *International Journal of Mass Emergencies & Disasters*, *13*(3), 221-229. <https://doi.org/10.1177/028072709501300301>

Rodrigues, F. H., & Abel, M. (2019). What to consider about events: A survey on the ontology of occurrents. *Applied ontology*, *14*(4), 343-378.

Romero, L. D., Padilla, J. J., Palacio, K., & Frydenlund, E. (2024). Mapping cooperation: insights into Colombia's humanitarian response to migration from Venezuela. *Frontiers in Human Dynamics*, *6*, 1345110.

Salem, S., Mansour, N. M., Hassan, I. H., & Debian, R. A. (2022). Investigating the Role of Conceptual Metaphor in the Representation of the Syrian Human Crisis in British Newspaper Articles: A Cognitive Approach. *Journal of Language and Literature*, *22*(1), 1-14.

Schröter, M., & Storjohann, P. (2015). Patterns of discourse semantics: A corpus-assisted study of financial crisis in British newspaper discourse in 2009. *Pragmatics and Society*, *6*(1), 43-66.

Shaikh, A., Mahoto, N. A., & Unar, M. A. (2019). A Lexicon based Approach Towards Concept Extraction. *3c Tecnología: glosas de innovación aplicadas a la pyme*, *8*(1), 50-67.

Srinivasan, V., & Kumar, P. (2015). Emergent and divergent resilience behavior in catastrophic shift systems. *Ecological Modelling*, *298*(C), 87-105. <https://doi.org/10.1016/j.ecolmodel.2013>. (Ecological Modelling)

Tainter, J. (1988). *The collapse of complex societies*. Cambridge university press.

*Understanding Collapse, in Stockholm Resilience Centre - Research.* (2018).

UNDRR. (2009). *UNISDR terminology on disaster risk reduction*.

Wallach, E. (2019). Inference from absence: the case of archaeology. *Palgrave Communications*, *5*(1), 1-10.

Wang, Y., Tian, J., Ones, D. S., & Landers, R. N. (2022). Using natural language processing and machine learning to replace human content coders. *Psychological Methods*.

Zouave, S. (2014). Manipulation in Newspaper Articles: A Political Discourse Analysis of Lexical Choice and Manipulation in Japanese Newspaper Crisis Reporting in the case of North Korea. In.

1. Comments

**Jose Comments**

**What is the main idea, and how will we accomplish it?**

* What are we going/attempting to do?
* What are the steps we propose to follow?
* What data sources are we going to use?
* How are we going to extract and analyze the data?
* What do we expect to find?

**Things to consider/expand on:**

* Balance: I have said this one too many times. Notice that GDELT has a paragraph allocated for explanation. Lexis Neis is two sentences, Meta is two lines, and Web Archive is four lines.
* The idea is not to use GDELT. Recall how complicated it is to analyze information from this place in a manner that results can be validated. The idea was to know how it extracts and categorizes news.
* How is TikTok useful for us considering that it would require video analysis? What does it provide?
* Overall, the data sources could elaborate more on how "good" or "reliable" they are for the intended purpose.
* Section 2: seems to have a bit more of thought. However, you mentioned lexicons. How many have we identified? Who has used lexicons that we can reference? Has cameo been used as a lexicon for the categorization of events?
* Section 3: how are we going to use metadata? What does it provide?

**Erika Comments**

**Part A: shock/not-shock 🡪 determination of “candidate” events**

1. GDELT data [brief description of *“what is GDELT”* + strengths + limitations w/citations]
   1. Variable 1 from GDELT + description + usage (strengths/limits)
      1. For example, tone – what it is; what it isn't; how to monitor events **(EVENTS NOT ARTICLES)** like w/change of means (by how much?)
      2. What GDELT labels are required?
   2. Variable 2 …
   3. Variable 3 …
2. Other datasets? How to round this out?
3. Testing strategy (sample + student rates accuracy (how??) + measure + decide if the shock/not-shock algorithm is sufficient)
4. Repeat (refine) until algorithms can identify (sufficiency/accuracy metrics?)

**Part B: Collapse/Not-Collapse (see Box 1 in Cumming & Petersen)**

1. Identity must be lost.
   1. This might be Wikipedia data ~ how?
   2. LexisNexis content analysis/topic modeling of the shift in infrastructure?
   3. Macrolevel indicators like the amount of conflict or change in a democracy?
2. Identity loss should happen fast (less than 1 generation of actors – not colloquial “fast”
   1. Maybe GDELT for news persistence?
   2. Surveys of case study countries? /Narratives of shift
   3. Doesn´t GDELT have a book ngrams search? Could that be useful?
3. Collapse involves substantial thresholds losses of social-ecological capital
   1. Maybe traditional global stats variables monitored by mean change like UN SDGs?
4. The consequences of collapse must be fasting
   1. Maybe Wikipedia data? ~ how?

**Extra comments: About second review**

1. **Data source section**
   1. In the data sources table (Table 1), I would rather see one column with a brief description of its use in the project, rather than Strengths/Limitations – I think you can maybe discuss those two in each data’s description?
   2. Add to each data description the strengths/ limitations.
   3. For google news, are you able to use the content of the articles, or is it restricted? What you’re missing in each of these data descriptions is “why” we should need it/how we will use it. I agree with Jose’s comment about reliability metrics (probably in the literature) for each dataset.
   4. To Jose’s comment about GDELT, you just need a brief 2-sentence description of it and why it’s good, but like Jose’s comment states (and like my comment above), concentrate that section more on the “why” and “how” of what we will use it for. You can give specific examples.
2. **Section 3: Phase 1**
   1. there is something wrong with the quote marks in the first sentence starting with “candidate”
3. **Section 3.1: Global Database**
   1. doesn’t sound specific enough. When you have Figure 2, I’m assuming that all the blue boxes are associated with the Lexicon development?
   2. The intro at 3.1 should tell the reader that specifically. Then it should tell the reader why you’re telling them these things.
4. **Section 3.1.1: Lexicon**
   1. GDELT uses CrisisLex – how are you going to use GDELT to extract more lexicons? That process isn’t clear to me.
   2. When you say “human-made disasters” in 3.1.1 (paragraph 1), what does that mean? Does that include financial/economic crisis?
   3. Overall: I find this section \*almost\* compelling, but right now it just feels really disjointed. I think you need to clean it up a bit and make it read as a recipe. It seems to rely on CrisisLex, but there are no other lexicons included/identified? Did you talk to Krzysztof, because he was also working on some lexicon stuff? I don’t like the figure right before 3.1.2
   4. I don’t think it adds. I think instead you would benefit from a picture that shows how a lexicon can be generated that complements what CrisisLex can offer (you say it’s only for natural and manmade disasters, but are there other limitations? What are the other lexicons for? How do you know when the collection is complete? How do you find those words? Look to how CrisisLex was made and maybe draw from that. This section just feels to me like it’s missing a lot of the “how” and “why” of the need for lexicons in this research.
5. **Section 3.1.2: News analysis**
   1. - Why “key individuals and their relationship with other entities”? – is this making this step more complicated than it needs to be?
   2. - I think I understand where you’re going with the Metadata part, but I think you need to look for a new way to explain it, because it’s just too confusing.
   3. - in Figure 3, why are the 3 final columns important? This isn’t clear in the narrative of the proposals.
   4. - web “scraping” not “scrapping”
   5. - I get confused about why the paragraph starting at the top of the 6th page begins with Google News API but suddenly ends with Web Archive – I don’t understand when each one is used.
   6. - “Content analysis is a method used in various disciplines to extract quantitative measures from textual information.” – this is not necessarily true. Lots of people do content analysis without any quantitative measures.
   7. - Explain why we need to find entities and relationships – this is not clear so I can’t understand why you need content analysis.
   8. I will be honest, most of the metadata part was lost on me. I think you need to rethink how you explain what, why, and how it is you’re doing what you’re doing.
   9. What does SNA give us in terms of understanding relationships? Why do we need this?
   10. Figure 4 doesn’t feel informative to me. I can’t visually see the clusters, and I don’t know what you’re using the clusters for.
6. **Section 3.1.3: Events clustering**
   1. You’ve completely lost me by 3.1.3 –
   2. What does this have to do with Figures 1 & 2?
   3. Why is this happening?
   4. I don’t know where you’re going with the figure right before 3.1.4 – how does this relate to Figures 1 & 2? There’s almost too many figures, but not quite the right ones that explain what I’m reading.
7. **Overall section 3.1.2 & 3.1.3**
   1. You have a tendency to not tell me what/why/how in each of your sections. You give me what, but it’s not linked to the how/why, which makes it hard to buy into.
   2. You’re losing the central “story” thread by changing the type of words you’re using over and over (messaging), and so you’ll need to clean up a bit the order that you say things in, but also tying up what you need to say before it gets too long.
   3. You really lost me by 3.1.3, so this is high priority to revisit and think how you will explain this to someone. Try to use Figures 1 & 2 as your grounding point.
   4. I will tell you what the Army Colonel who I once worked for (he was actually originally someone very high up in Microsoft) told me: whatever you decided to write, take out all the adjectives. That applies in this case – and ChatGPT is notorious for adding too many adjectives. In section 3.1.1, consider removing at least half of the adjectives and adverbs.
8. Extra: News Database

A group of colorful circles

Description automatically generated

Figure 6. Graph representation of the final database

1. <https://sociology.osu.edu/worldhandbook> [↑](#footnote-ref-2)
2. <https://www.quantexa.com/> [↑](#footnote-ref-3)
3. <https://www.textmap.com/> [↑](#footnote-ref-4)
4. https://oduwsdl.github.io/hypercane/ [↑](#footnote-ref-5)
5. https://atlas.nomic.ai/ [↑](#footnote-ref-6)
6. https://pypi.org/project/gnews/ [↑](#footnote-ref-7)
7. https://www.lexisnexis.com/en-us/professional/academic/nexis-uni.page [↑](#footnote-ref-8)
8. https://socialmediaarchive.org/?ln=en [↑](#footnote-ref-9)
9. https://transparency.fb.com/researchtools/meta-content-library [↑](#footnote-ref-10)
10. https://fort.fb.com/researcher-datasets [↑](#footnote-ref-11)
11. https://archive.org/ [↑](#footnote-ref-12)
12. https://www.wikidata.org/wiki/Wikidata:Main\_Page [↑](#footnote-ref-13)