PROCEEDINGS OF SCIENCE

PoS(ISGC2024)009

A Lexicon for Social Media-Based Cultural Heritage Information in Crisis Situations: A Proposal

Anastasiya Sopyryaeva,[a] **Elisabetta Ronchieri,**[a,b,*] **Ahmad Alkhansa,**[a]
Alessandro Costantini[a] **and Alessandro Bombini**[c]

[a] *INFN CNAF*
Viale Berti Pichat 6/2, Bologna, Italy

[b] *Department of Statistical Sciences, University of Bologna,*
Via Belle Arti, 41, Bologna, Italy

[c] *INFN Sezione di Firenze*
, Firenze, Italy

E-mail: anastasiya.sopyryaeva@gmail.com, elisabetta.ronchieri@cnaf.infn.it, ahmad.alkhansa@cnaf.infn.it, alessandro.costantini@cnaf.infn.it, bombini@infn.it

[*] Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

Social media can play a crucial role in disseminating information about cultural heritage if a proper lexicon is available and able to identify valuable data for the management of crises that are caused by either natural or human-induced disasters. The lack of published studies concerning terminological resources for cultural heritage (neither generally, nor in the context of social media discussion) and the absence of a lexicon dedicated to detecting cultural heritage-related tweets on social media during crisis events have driven us to investigate such an area of research.

For such reason, we have undertaken the task of creating our lexicon that provides essential information, comprehends the domain, and facilitates further research in the field. The lexicon has been defined according to keywords that are commonly used on social media for a specific discussion, and are represented in a list of uni-gram and bi-gram terms from natural language processing solutions: e.g., culture or ancient site are keywords for cultural heritage discussion, while vandal or property damage are keywords for vandalism discussion. Furthermore, the defined lexicon can be representative of the domain but also accurately reflect the specific vocabulary commonly utilized within social media platforms, such as Twitter.

Developing a representative lexicon is an essential preliminary step in this study because we have to devise a method for identifying Twitter messages that are related to the field of cultural heritage management in crises. The raw datasets have been collected from January 1 to April 27, 2023, with the Twitter API, in the context of the 4CH project (European Competence Centre for the Conservation of Cultural Heritage) that aims at setting up the methodological, procedural, and organizational framework of a Competence Centre able to seamlessly work with a network of national, regional, and local cultural institutions.

Our dataset is extensive and originates from diverse time periods, events, and geographical locations. These distinct locations encompass various nations and institutions, each with its distinct interpretations and definitions of culture and its elements. Questions regarding the nature of culture and what constitutes heritage lack general clear answers on an international scope. Given this complexity, we have chosen to create a lexicon that provides the most general framework as possible, relying on the documents of The United Nations Educational, Scientific and Cultural Organization that include vocabularies close to those we intend to create for cultural heritage.

1. Background

There is ongoing interest in the investigation of the role and impact of social media platforms on cultural heritage sustainability and cultural preservation. Social media research for crisis management is an established domain. Literature reviews within this field [1, 2] reveal a variety of issues faced by social media researchers in the crisis management domain. However, the majority of studies concentrate on emergencies affecting human health and safety [3]. There are many works devoted to people's emotions [4, 5] or classification of messages based on people's needs during disaster events [6, 7]. Instead, this research shifts the focus toward crises' impact on cultural heritage, which has not been extensively explored in the academic literature.

In this study, we have explored datasets sourced from the Twitter API to uncover valuable information for cultural heritage management in crises that are triggered by either natural or human-induced disasters. Furthermore, we have developed methodologies that leverage artificial intelligence (AI) technologies to efficiently extract this valuable information from the datasets.

To facilitate the textual data exploration, we have created lexicons able to detect tweets that hold relevance for the cultural heritage domain during crisis events, extract essential information embedded within these tweets, comprehend the domain, and facilitate further research in the field. Therefore, we have created a lexicon of keywords related to the cultural heritage domain and vandalism. The keywords should be represented in a list of uni-gram and bi-gram terms that are commonly used in the context of cultural heritage or vandalism discussions on social media. To provide an example, terms like culture or *ancient site* can be added as keywords for cultural heritage discussion, while terms like *vandal* or *property damage* can be added as keywords for vandalism discussion.

These lexicons should not only be generally representative of the domains but should also accurately reflect the specific vocabulary that is commonly used within social media platforms, like Twitter. The importance of domain-specific lexicon has been introduced by computational social scientists, highlighting the semantic and sentiment diversity of natural language [8, 9]. To give an example, by examining the vocabulary of the tweet corpus related to vandalism, we have obtained frequent terms related to sentiment expression of hate and anger, such as *hate crime*, *sexual abuse*, *ethnic genocide*. Those terms are not explicitly associated with the vandalism domain, however, they are widely used by people on social media and we have decided to include them in the defined lexicon.

Developing representative lexicons is an essential preliminary step in achieving our research objectives because we aim at devising a method for identifying Twitter messages that are related to the field of cultural heritage management in crises. The raw datasets obtained from the Twitter API, despite being collected based on keywords, contain numerous irrelevant tweets and are not suitable for investigation within the context of cultural heritage management in crises. Additionally, these lexicons can enhance the utility of machine-learning classification algorithms by serving as a reference point for manual labeling and semi-supervised classification techniques. Consequently, they can be applied to other similar datasets of tweets.

PoS(ISGC2024)009

2. Text Preparation

Before starting with the text preparation and data exploration process of relevant tweets to the cultural heritage management in crises, it is essential to introduce the datasets taken into consideration. We have collected 13 datasets from Twitter API by considering 13 different categories in the period from January 1 to April 26, 2023: 11 categories related to different disaster events, particularly bombing, downpour, earthquake, explosion, fire, flood, hail, landslide, squall, tsunami, and volcano; 1 category named vandalism; 1 category named cultural heritage. Each query used to retrieve data from Twitter API has been composed as follows:

```
category name lang:en  -is:retweet  -is:reply
```

The initial stage in any Natural Language Processing (NLP) pipeline involves text normalisation, where the raw text must be preprocessed before any statistical modelling. Below there are the fundamental normalisation procedures implemented on raw tweets within this research:

- Remove hashtags, emojis, and URLs from tweets - They are unnecessary for the text models implemented in this research; nevertheless, hashtags, emojis, and URLs carry valuable information for the investigation of the domain, thus they are separated as textual features and stored as distinct text attributes.

- Remove HTML markup, special characters, punctuation marks, whitespace, mentions (@mention), retweet identification (RT);

- Remove or replace numbers;

- Lowercase all characters;

- Remove stopwords - The choice of a stopword list depends on task objectives. The three stopword lists used in this study are: NLTK English stopwords that contain words commonly found in English texts as a whole (e.g., *do* and *have*); domain-specific stopwords that contain primary keywords aligned with dataset categories, which serve as the basis for the initial collection of tweets via Twitter API, such as bombing, downpour, earthquake, explosion, fire, flood, hail, heritage, landslide, squall, tsunami, volcano, vandalism. These words appear in every tweet and thus, they are dispensable for the majority of tasks; custom stopwords that include common English words that are absent from the NLTK collection, such as *im*, *iam*, *dont thats*, *didnt*, *hes*, *doesnt*, *youre*, *theyre*, *heres*, *theres*, *like*, *youll*, *oh*, *ive*, *yo*, *yall*, *vs*, *really*, *isnt*, *waaaah weewoo*, *weee*, *th*, *rd*.

- Implement lemmatization or stemming - Lemmatization is preferred over stemming when morphological information about the words is essential. For example, stemming the inflected words *historical* and *history* leads to the same stem *histori*, while lemmatizing them gets two lemmas *historical* and *history* as the part of speech of a word is preserved. For most of the tasks in this study, we have chosen lemmatization.

  Below there is an example of a tweet before and after the normalisation procedures, replacing the real name of a football player with [*Name*]:

PoS(ISGC2024)009

```
Name is consistently BOMBING footballs.\n\nAll. Year. Long.
```

Example of a tweet after normalisation:

```
name consistently football year long
```

Following the text normalisation procedures, we have conducted basic word frequency investigation, such as n-gram analysis and word cloud visualisation. This delves into the vocabulary of the dataset, pinpoints opportunities for feature extraction, and facilitates the detection of tweets. The objective of this exploration has been to discover potential keywords that might aid in the effective identification and classification of relevant tweets.

The text preparation pipeline unfolds as follows: load text column containing raw tweet text for a specified category; remove duplicates; filter out irrelevant tweets, discarding those containing words irrelevant to the domain; implement text normalization; store cleaned text, extract features, and record the dataset reduction throughout each procedure.

3. Irrelevant Tweets Removal

Initial text exploration not only serves to define a collection of *relevant* domain keywords but also to discern words *unrelated* to the domain. Certain category keywords used for gathering tweets from the Twitter API possess complex semantics, and they also contain numerous irrelevant tweets. For instance, category keywords, such as *fire* and *landslide*, are the most ambiguous ones due to their multiple meanings and informal usage within the Twitter community.

More precisely, in politics, the term *landslide* refers to a victory in which the candidate receives an overwhelming majority of the votes [1]. Consequently, it turns out that around 50% of the *landslide* tweets refer to presidential elections in the USA. For example: "Belarus election: Name's claim of landslide victory sparks widespread protests (from The Guardian)". Similarly, the term *fire*, apart from its primary connotation, is a slang term for awesome or cool [2]. Consequently, roughly 50% of the *fire* tweets refer to music, videos, computer games, shopping items, or the physical appearance of individuals. For example: "This dress is perfect on you! Looking FIRE!!".

Category	**Irrelevant terms**
landslide	fleetwood mac, win, won, winner, trump, victory, election, vote, elon musk, presidenty, winning, poll, reelection, biden, republican, democrat, candidate, voter, song, donald, lgbt, contraception, abortion, migrant, polling, voter, elected, voting
fire	game, games, gaming, gamer, player, play, playing, played, tv, video, music, radio, ass, song, sing, amazon
heritage	ebay, baseball, football
downpour	kpop
tsunami	lgbtq

Table 1: Lists of irrelevant terms per category.

To enhance the informativeness of ambiguous datasets and minimize their size, we have defined queries of irrelevant words for each dataset and removed tweets that contain these words. These

[1] Grammarist. "Landslide". (accessed 05/06/2024). https://grammarist.com/idiom/landslide/

[2] Slang.net. "Fire". (accessed 05/06/2024). URL: https://slang.net/meaning/fire

PoS(ISGC2024)009

irrelevant words have been manually collected from uni-gram analysis of each category corpus. A word has been added to a list of irrelevant terms to a specific category, only if it refers to another category. For example, a term *voting* has been added into a list of irrelevant terms to the landslide category, while a term *gamer* has been added as irrelevant to the fire one. Table 1 lists these irrelevant words for the dataset categories where this approach has been applied. Some categories do not have irrelevant words lists since there are no irrelevant terms found during uni-grams investigation. After applying the procedure that removes tweets with irrelevant words, we have obtained the following statistics showing a decrease in dataset size, as shown in Table 2.

Category	**Initial Tweet Count**	**After Duplicates Removed Count**	**After Irrelevant Tweets Removed Count**	**Final Tweet Count**
downpour	22,257	20,085	19,838	19,838
bombing	125,919	121,032	-	121,032
earthquake	1,084,887	1,038,844	-	1,038,844
explosion	272,393	253,553	-	253,553
fire	5,487,850	4,949,689	3,989,307	3,989,307
hail	528,962	497,495	-	497,495
heritage	489,089	467,244	441,436	441,436
landslide	67,842	61,591	28,049	28,049
flood	626,392	586,986	-	586,986
squall	19,189	16,541	-	16,541
tsunami	69,783	64,945	64,241	64,241
volcano	104,521	94,991	94,301	94,301
vandalism	44,503	43,131	-	43,131

Table 2: The absolute number of data entries remained after each dataset cleaning phase.

Another source of ambiguity identified during the vocabulary exploration involves abbreviations of both informal slang and formal abbreviations. For example, a token *SAN* can refer to the abbreviation of the USA State *San Diego*, *saint* or any country/city beginning with *San*, such as San Marino or San Antonio. A token *sh* is a slang abbreviation that can refer to *Same Here*, *Self-Harm*, or *Sexual Harassment*. The frequency of such ambiguous words is high - many abbreviation terms are found in the top 200 uni-gram terms of the corpus. This frequency and, thus, complexity of our vocabulary arises from several factors. Firstly, Twitter is a social community with unique slang, which requires special expertise during the analysis process to disambiguate these tokens. Secondly, the collection of tweets spans diverse locations and events, resulting in the inclusion of highly localised abbreviations that can be challenging to decipher. Thirdly, these abbreviations might carry multiple interpretations, further complicating their understanding. In the present stage of the research, we retain all unfamiliar abbreviations, as they could potentially hold relevance within the category.

4. Hashtags and Emojis

On social media, people label their messages with hashtags to highlight the content, or the key events/places/people/phenomenon, of the messages. Our exploration of hashtags in each category dataset, indeed, shows that the authors have used hashtags to highlight the message content. Interestingly, there are hashtags that belong to irrelevant terms. These tweets contain category keywords, such as *bombing*, but the hashtags tell us that the keyword is probably irrelevant to our category. To give an example, the following messages belong to the *bombing* dataset and have been retrieved with the keyword *bombing*:

Tweet 1	He's a BIG BOI!! Lmao look at that lil Chespin photo bombing Quilladin #PokemonGO #PokemonGOCommunityDay #Chespin
Tweet 2	Loved this look back at the 2013 Name by @Name in @TheAthletic, gave me goosebumps thinking about what that team meant to the city in the wake of the Marathon Bombing. An amazing team, an amazing piece. #Name

The texts in the two tweets contain the word *bombing*, however, they do not relate to the *bombing* event. They both relate to games, either online or offline, in which the context of the word *bombing* is often used. Hashtags help us to understand that the messages are used in the context of games, which is irrelevant to the disaster category. In the first tweet the hashtags refer to a famous online game PokemonGo (#PokemonGO #PokemonGOCommunityDay), while in the second tweet the hashtags are referred to a name of a famous team [*Name*] (#Name). We can consider this observation when labelling tweets as relevant to the domain in question. Table 3 represents several illustrations of a possible hashtag based labelling.

Tweet	**Hashtag**	**Label**
The coalition of America, Saudi Arabia, UAE and of course with the help of Israel and England have been attacking ... and besides bombing the ..., they are killing them with starvation and severe siege. #BlockadeIsWar	#BlockadeIsWar	bombing_related
A father who lost his daughter younger than 8 yrs and a mother who lost her son to .... Sobbing mother bewailed “nothing can console" her grief. #WhatsHappeningInMyanmar	#WhatsHappeningInMyanmar	bombing_related

Table 3: Hashtag based labeling examples.

This approach can be rather useful in the exploratory analysis or as a semisupervised labelling technique. In order to use this approach for automatic labeling, we need to have either a comprehensive list of all topics in which a keyword bombing can be used to exclude irrelevant tweets, or a list of all possible hashtags that are only used in a bombing-disaster context. This is not a reasonable approach. Instead, we can apply hashtag-based filtering on the stage of exploratory analysis to filter out irrelevant tweets and obtain a list of tweets labeled as relevant disasters with high confidence to use during the usage of machine learning (ML) classification algorithms. This approach has already been used by social media researchers. The literature review in the study by Qiang [10] reveals that hashtags have been used in various studies in order to facilitate the correct topic discovery of tweets. In addition, inspecting the common hashtags used in our tweets, we have observed that hashtags can be used to facilitate tweet topic detection (#WhatsHappeningInMyanmar); identify locations the tweet is talking about (#gaza); identify sentiments, and calls to action (#weremember); identify irrelevant tweets (#PokemonGO).

Emojis are highly ambiguous text features, mostly because it is uneasy to interpret users' meaning when they use a particular emoji: especially when it is used sarcastically, and when it is

PoS(ISGC2024)009

a real emotion. However, emojis exploration before applying ML models is useful to get a clue on what kind of sentiment analysis can be applied to the specific dataset.

5. Domain Lexicons for Tweets Detection

Considering the insights gained from the initial text exploration, it is apparent that constructing accurate domain lexicons is critical for effective tweet detection. Given that our datasets are not labeled by relevance and manual labeling would be too resource-intensive, the application of ML classification algorithms is not feasible in the current research stage. However, future considerations could include manual labelling, transfer learning, weak supervision, or unsupervised labelling methods, all of which require substantial domain expertise and a comprehensive seed set of representative tweets. Consequently, our work is focused on keyword-based tweet detection algorithms. The main input component of such an algorithm is a keyword list, which serves as a reference to identify tweets aligned with the domain. The objective of this section is to introduce the lexicons of keywords defined for the domains under investigation. These lexicons are pivotal components to accurately detect tweets that encapsulate the essence of the respective domains.

Adhering to the categorization of our collected datasets into disasters, vandalism, and cultural heritage, we have constructed three corresponding lexicons. These lexicons can be combined and refined to yield a single lexicon specifically tailored for cultural heritage in crisis scenarios. Typically, the approach to lexicon generation involves 2 main steps: the initial query generation step, followed by the query expansion step [11]. In the following, details for the disaster and cultural heritage lexicons are provided.

5.1 Disaster Lexicon

The keywords-based approach has been extensively utilised by researchers in the detection of disaster-relevant tweets, leading to the creation of several valuable resources, such as CrisisLexRec [12] and EMTerms (Emergency Management Terms) [6] that represent the most cited, utilised, and comprehensive lexicons. We have adapted these resources for the application to our specific case study. The creators of these lexicons also offer collections of tweets, labelled or not, as well as a tool for constructing diverse lexicons based on their established methodologies, aiming to facilitate research in this field.

CrisisLexRec is a crisis lexicon developed upon the CrisisLexT6 collection of tweets that includes English tweets across six large events in 2012 and 2013, with about 10,000 tweets labeled by relatedness (as "on-topic", or "off-topic") with each event. The overall number of tweets is 60,000. This collection encompasses a range of crises, including events like the 2012 Sandy Hurricane, 2013 Boston Bombings, 2013 Oklahoma Tornado, 2013 West Texas Explosion, 2013 Alberta Floods, and 2013 Queensland Floods. The corresponding CrisisLexRec lexicon has been initially dedicated to sample messages related to crises across a variety of crisis events. Comprising 380 terms, mainly in bigram format (e.g. *flood crisis*, *bombing suspect*, and *victims*), CrisisLexRec stands as a valuable resource for our purposes of relevant tweet detection.

EMTerms is another open-source lexicon developed by the research community. Concerning CrisisLexRec, it embraces diverse terminology. EMTerms consists of 7,200 terms categorised into 23 information type categories, in contrast to 380 general terms in CrisisLexRec. EMTerms aims to create a lexicon, which reflects the real, observed linguistic expressions used in Twitter to describe a wide-range variety of crises, and, at the same time, focuses on the information needs of emergency managers. This lexicon resource considers terminology commonly used within crisis-related tweets, fully represents the vocabulary used by people in social media discussions of disasters, and provides an extensive categorisation of these terms. EMTerms represents a comprehensive resource, however, it can not be used for tweets retrieval and collection, because the terminology presented there is not only used within crisis discussions but also in many other cases. Consider the term *airport terminal still closed*, while being a term commonly used within crises, this term may be also used in day-to-day life, referring to the closure of an airport at nights or due to technical issues. Applying such terms to tweet detection is not useful as it will retrieve many non-disaster-related tweets. Therefore, EMTerms is too extensive for our task of tweet detection. EMTerms should be applied on tweets that have been already classified as relevant to disaster.

To illustrate the claims about the usage of the lexicons in our research, we have tried to apply both lexicons to detect disaster-related tweets in our specific datasets. CrisisLexRec has been developed around five types of events, only three of which align with our datasets. While general disaster-related terms, such as *death* and *injury*, can be applied to any disaster event, the exclusive use of CrisisLexRec for our 13 datasets supports us in the omission of event-specific keywords and, subsequently, the tweets. This is particularly noticeable for event types that are not part of CrisisLexRec's development scope. EMTerms, due to its extensive nature, indeed appears to be overly broad for direct application as a basis for tweet detection.

Taking into account the obstacles to the usage of open-source crisis lexicons, we have decided to construct our lexicon. Following the two steps to generate a lexicon, our approach to disaster lexicon generation involves the initial query generation step and the query expansion step. Figure 1 shows the characteristics of each step to generate disaster lexicons.

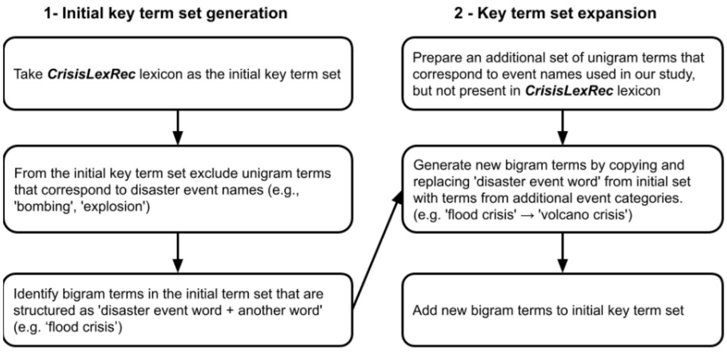
1- Initial key term set generation
2 - Key term set expansion
Take ***CrisisLexRec*** lexicon as the initial key term set
From the initial key term set exclude unigram terms that correspond to disaster event names (e.g., 'bombing', 'explosion')
Identify bigram terms in the initial term set that are structured as 'disaster event word + another word' (e.g. 'flood crisis')
Prepare an additional set of unigram terms that correspond to event names used in our study, but not present in ***CrisisLexRec*** lexicon
Generate new bigram terms by copying and replacing 'disaster event word' from initial set with terms from additional event categories. (e.g. 'flood crisis' → 'volcano crisis')
Add new bigram terms to initial key term set

Figure 1: Lexicon generation workflow for disaster domain.

Examining the initial key term set generation, for our specific tweet detection objective, the CrisisLexRec lexicon has been chosen as the most fitting option to serve as the initial disaster key term set. It contains terms that are relevant, precise, and explicit and thus can be used to retrieve tweets. To address the challenge of CrisisLexRec being built upon the limited number of event types, we have decided to enhance the CrisisLexRec lexicon by incorporating key terms that are specific to the other types of events present in our datasets. This strategy allows us to bridge the

PoS(ISGC2024)009

gap and ensure that our lexicon covers a broader range of event types and related terms. Before proceeding to the expansion, we examine the CrisisLexRec in detail to understand its nature and expand the lexicon accordingly: refinement of CrisisLexRec by manually examining the terms and excluding uni-grams from the CrisisLexRec lexicon that correspond to disaster event names (e.g. *bombing*, and *explosion*). These terms tend to be too general and can lead to semantic ambiguity being used as single key terms; identifying relevant bi-grams by selecting the bi-grams present in the CrisisLexRec lexicon, particularly those structured as disaster event word (e.g. *explosion*) and another word (e.g. *crisis*). These bi-grams have the potential to be adapted to our specific event categories.

Examining the key term set expansion, the chosen approach for lexicon expansion involves the following activities: creating additional bi-grams to include as key terms to the lexicon by replacing the *disaster* event word in the original bi-grams with the corresponding terms from our event categories (e.g. *explosion crisis* ← *volcano crisis*); adding new terms that are generated in the third phase of the initial term set.

5.2 Cultural Heritage Lexicon

The task of creating a set of keywords that is discriminative within the cultural heritage domain requires a deep understanding of the field. To establish this knowledge base, a literature review has been conducted, encompassing existing attempts to define terminology within the cultural heritage domain [3, 13, 14]. A review of the role of social media for the cultural heritage sustainability [13] reveals that there is ongoing interest in the investigation of the role and impact of social media platforms on cultural heritage sustainability and cultural preservation. However, there is not much research found concerning terminological resources for cultural heritage neither generally, nor in the context of social media discussion. The papers investigating natural language processing within the cultural heritage management domain mostly focus on explaining the high ambiguity of cultural heritage terminology, its diversity across locations [14] and the application of control vocabularies approach for cultural heritage management [14].

In the absence of a lexicon dedicated to detecting cultural heritage-related tweets on social media, we have undertaken the task of creating our lexicon for this purpose. Our approach to lexicon definition has been based on the following premises. Our dataset is extensive and originates from diverse periods, events, and geographical locations. These distinct locations encompass various nations and institutions, each with their distinct interpretations and definitions of culture and its elements. Questions regarding the nature of culture and what constitutes heritage lack general clear answers on an international scope. In addition, we have taken into account that the texts collected are in English. This implies that users either come from English-speaking countries or, if they come from other regions, communicate in English due to their connection with an international community or a desire to address global issues using an international language. Because of this complexity, we have chosen to create a lexicon that provides the most general possible framework, relying on the documents of the United Nations Educational, Scientific and Cultural Organization (UNESCO) [3]. UNESCO focuses on the heritage found across the globe, thus, encompassing diversities of locations and nations, uniting them at the same time. The vocabulary

[3] UNESCO in brief". (accessed 5/6/2024). https://www.unesco.org/en/brief

used in the UNESCO documents, therefore, is assumed to be the closest to the vocabulary of the custom lexicon for the cultural heritage we aim to create. By rooting our lexicon to the universal principles of UNESCO, we have created a lexicon that captures the essence of cultural heritage discussions across diverse locations, nations, and time periods.

Figure 2 shows the characteristics of each step to generate a cultural heritage (CH) lexicon.

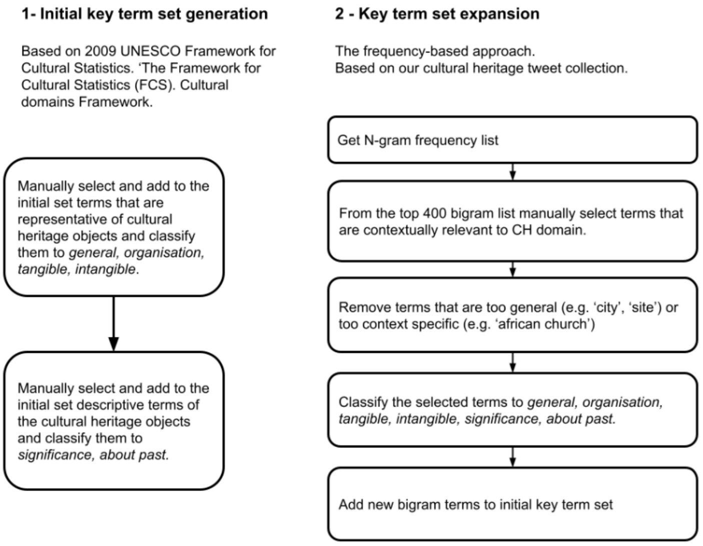
1- Initial key term set generation
Based on 2009 UNESCO Framework for Cultural Statistics. 'The Framework for Cultural Statistics (FCS). Cultural domains Framework.
Manually select and add to the initial set terms that are representative of cultural heritage objects and classify them to *general, organisation, tangible, intangible*.
Manually select and add to the initial set descriptive terms of the cultural heritage objects and classify them to *significance, about past*.
2 - Key term set expansion
The frequency-based approach.
Based on our cultural heritage tweet collection.
Get N-gram frequency list
From the top 400 bigram list manually select terms that are contextually relevant to CH domain.
Remove terms that are too general (e.g. 'city', 'site') or too context specific (e.g. 'african church')
Classify the selected terms to *general, organisation, tangible, intangible, significance, about past.*
Add new bigram terms to initial key term set

Figure 2: Lexicon generation workflow for CH domain.

Describing the initial keyword set generation and considering the global scope of this study, we have demanded the most universal and neutral framework for CH definition. This is why our attention has turned to UNESCO, the largest international organization with an extensive history of involvement in global culture and preservation-related matters. UNESCO promotes "international cooperation in education, sciences, culture, communication, and information". UNESCO's engagement in worldwide culture and heritage issues, spanning decades, makes it a reliable source of domain expertise. Defining culture, cultural heritage, and its keywords, we particularly refer to UNESCO documents. By intensifying interconnection, and interchange between nations, UNESCO has developed and released a series of documents aimed at presenting frameworks in the cultural heritage field. One document is the 2009 UNESCO Framework for Cultural Statistics [4] that provides the measurement of a wide range of cultural expressions. Before defining the set of CH-related key terms, it is crucial to clarify what is understood by CH domain. UNESCO includes artifacts, monuments, a group of buildings and sites, and museums that have a diversity of values including symbolic, historic, artistic, aesthetic, ethnological or anthropological, scientific, and social significance. It includes tangible heritage (movable, immobile, and underwater), and intangible cultural heritage (ICH) embedded into cultural, and natural heritage artifacts, sites, or monuments. The definition excludes ICH related to other cultural domains such as festivals and celebrations. It covers industrial heritage and cave paintings [5].

[4] The 2009 UNESCO Framework for cultural statistics (FCS), 2009.

[5] Cultural heritage. Definition. (accessed 5/5/2024). https://uis.unesco.org/en/glossary-term/

PoS(ISGC2024)009

This broad definition can be discerned into two primary focuses: the objects that can be considered as CH (e.g., buildings, sites, and museums) and descriptive words that help to identify if the objects belong to cultural heritages or not in particular contexts (e.g., historic, artistic, and anthropological). The UNESCO Framework for Cultural Statistics dives deeper into each of the cultural heritage objects. We have decided to extract the initial seed of keywords relying on this framework. Here, UNESCO defines, categorizes, and describes cultural domains and related domains thus being a valuable resource for defining keywords.

Relying on the UNESCO framework for cultural statistics domains, we have first included in the list of keywords all the object words that correspond to cultural heritage objects within each domain (e.g., music, fine arts, and manuscript). Second, we have manually read all the definitions of those objects, and using domain understanding, we have chosen terms that are often repeated and are important in the manner that they are descriptive words of CH phenomena itself and of the objects (e.g., traditional, historical, and cultural). These descriptive words are considered to be helpful because they point to the cultural significance of the object they describe. Lastly, we have finalised the initial CH keywords set. It contains 70 terms that we have labelled manually with respect to 6 categories, namely: general that refers to the most generic and definitive terms for cultural heritage domain (e.g., *heritage*); organisation representing any organisation/institution that may hold cultural significance (e.g., *unesco*); tangible including terms that represent tangible cultural heritage objects categorised by UNESCO (e.g., *musical instrument*); intangible that refers to terms representing intangible cultural heritage objects (categorised by UNESCO) (e.g., *ritual*); significance including descriptive terms expressing cultural significance/importance of a phenomenon (e.g., *symbolic*); about past including descriptive terms expressing the historical significance of a phenomenon (e.g., *ancient*). Before proceeding to the lexicon expansion step, we have investigated how well our initial lexicon is representative of the domain. We have checked how often terms in our lexicon appear in the CH tweet collection. We have created lists of uni-gram and bi-gram terms' frequencies in the CH dataset of tweets and calculated the intersection of lexicon terms with the terms in frequency lists. The result revealed 62 matches out of 70 terms in the initial keyword set: 56 within uni-grams and 6 matches with bi-grams.

The query expansion step is crucial for refining the initial lexicon extracted from UNESCO documents, as these terms may be too formal or too general to capture the specific semantics used by the Twitter community. During the query expansion step, we have used a frequency-based approach. The idea is to investigate the CH dataset of tweets to consider the social media vocabulary. We have relied on term frequencies within this dataset to decide the relevance and importance of new terms within the domain. The query expansion process is characterized by manually performing n-gram frequency analyses on the CH dataset to select the terms that are commonly used within the context of CH discussions on Twitter. These insights help identify frequently used and contextually relevant terms. From the top 400 bi-gram list we have manually selected terms that are contextually relevant to the CH domain. The preference for bi-grams over uni-grams is grounded in the aim to capture more semantics and reduce ambiguity. After experimenting with the lexicons and tweets, we have discovered that uni-grams often introduce noise due to semantic ambiguity, whereas bi-grams provide additional context that aids in disambiguation and detection of more discriminated tweets.

PoS(ISGC2024)009

cultural-heritage

While this approach may result in the loss of potentially valuable tweets, it is a necessary step to create a useful tweet collection for further exploration. This paper primarily focuses on exploratory analysis, seeking to uncover the diverse information that emerges within disaster-related tweets discussing cultural heritage. The intention is to better understand the range of topics before evaluating the approach or lexicon, especially given the present stage of the research.

During this step, the following terms have been added to the lexicon: terms falling into the category *significance* which express national and cultural identity (e.g., *national identity*, *national music*); terms corresponding to cultural heritage objects belonging to tangible' and 'intangible' categories (e.g., *church*, *temple*, and *city*); synonyms to the initial key terms (e.g., *heritage list* as a synonym to *world list*). The curation step is performed by filtering out general (e.g., *city*, and *site*) or context-specific (e.g., *African church*) terms from the list. New selected terms are labeled within 6 categories, called general, organisation, tangible, intangible, significance, and about past; later they have been incorporated into the lexicon. The length of the final lexicon is 84 terms as summarized in Table 4.

5.3 Vandalism Lexicon

In the absence of a lexicon dedicated to detecting vandalism-related tweets on social media, we have undertaken the task of creating our own lexicon for this purpose. The approach taken to construct the lexicon for vandalism is similar to what we have done for cultural heritage. It has involved a two-fold process: first, deriving a seed set of words from domain-specific literature [15, 16] and then conducting a frequency analysis of the dataset to identify terms that are actively used and can be incorporated into the lexicon. Williams [15] conducts a sociological survey that identifies forms of vandalism actions that can be used when adapting lexicon to social media context. Chatzigiannis [16] discusses sociopolitical and aesthetic aspects of the vandalism phenomenon from the conservator's point of view, providing vandalism definitions.

For the initial key terms set, we have created the list of terms identified by Chatzigiannis [16] (e.g., arson, break, and destruct). We have also read all the definitions of vandalism actions, and chose terms descriptive of vandalism acts (e.g., hostile and harmful). Furthermore, we have finalised the initial vandalism keywords set. It contains 38 terms. We have categorised each term by labeling them with a manually defined keyword category, namely: *general* that includes the most generic terms for vandalism domain (e.g., *crime*); *theft* (e.g., *looting*); *illegal* (e.g., *willful*); *conflict* (e.g., *terrorist*); *act against property* (e.g., *graffiti*).

Before proceeding to the lexicon expansion step, analogously with the CH lexicon, we have investigated how well our initial lexicon is representative of the domain. We have checked how often terms in our lexicon appear in the Vandalism tweet collection. We have created lists of uni-gram and bi-gram terms' frequencies in the vandalism dataset of tweets and calculated the intersection of lexicon terms with the terms in frequency lists. The result revealed 37 matches out of 38 terms in the initial keyword set: 36 within uni-grams and 1 with bi-grams.

During the query expansion approach, we have used frequency-based approach. The idea is to investigate the vandalism dataset of tweets to consider the social media vocabulary. We have relied on term frequencies in order to decide the relevance and importance of new terms within the domain. During the query expansion process, we have conducted n-gram frequency analysis on the vandalism dataset. From the top 400 bi-gram list we have selected terms that are contextually

PoS(ISGC2024)009

Term	Category Name	Source	Term	Category Name	Source
culture	general	initial	art	general	initial
heritage	general	initial	history	general	initial
religion	general	initial	world site	general	initial
world list	general	initial	unesco	organisation	initial
united nations educational, scientific and cultural organization	organisation	initial	museum	organisation	initial
gallery	organisation	initial	library	organisation	initial
academy	organisation	initial	religious	significance	initial
cultural	significance	initial	artistic	significance	initial
historical	significance	initial	spiritual	significance	initial
symbolic	significance	initial	ethnological	significance	initial
anthropological	significance	initial	archaeological	significance	initial
traditional	about past	initial	inherited	about past	initial
artisanal	about past	initial	ancient	about past	initial
ethnic	about past	initial	ancestry	about past	initial
patrimony	about past	initial	artefact	tangible	initial
artifact	tangible	initial	monument	tangible	initial
building	tangible	initial	architecture	tangible	initial
archive	tangible	initial	book	tangible	initial
manuscript	tangible	initial	literature	tangible	initial
work of art	tangible	initial	collection	tangible	initial
fine arts	tangible	initial	visual arts	tangible	initial
sculpture	tangible	initial	painting	tangible	initial
drawing	tangible	initial	handicrafts	tangible	initial
musical instrument	tangible	initial	clothing item	tangible	initial
jewellery	tangible	initial	decoration	tangible	initial
funerary	tangible	initial	traditional practice	intangible	initial
traditional lifestyle	intangible	initial	cuisine	intangible	initial
value system	intangible	initial	belief	intangible	initial
festive event	intangible	initial	sacred ceremony	intangible	initial
ritual	intangible	initial	celebration	intangible	initial
performing arts	intangible	initial	theatre	intangible	initial
dance	intangible	initial	opera	intangible	initial
puppetry	intangible	initial	festival	intangible	initial
fest	intangible	initial	fair	intangible	initial
intangible	intangible	initial	tangible	tangible	initial
world heritage	general	expanded	heritage site	general	expanded
future generation	general	expanded	destruction cultural	significance	expanded
amp heritage	general	expanded	heritage list	general	expanded
rich heritage	general	expanded	heritage significance	significance	expanded
natural heritage	general	expanded	national heritage	general	expanded
national music	intangible	expanded	national identity	significance	expanded
temple	tangible	expanded	church	tangible	expanded

Table 4: CH lexicon terms.

relevant to the vandalism domain. During this step, the following terms have been added to the lexicon: terms falling into the category *act against property* which have not been discovered before (e.g., *infrastructure invade*); synonyms to the initial key terms (e.g., *scandal* as a synonym to *conflict*).

The new selected terms have been labeled with 5 categories (i.e., general, theft, conflict, act against property, illegal) and incorporated into the lexicon. The length of the final lexicon is 48 terms as shown in Table 5.

Term	Category Name	Source	Term	Category Name	Source
crime	general	initial	vandalism	general	initial
vandal	general	initial	assault	general	initial
harmful	general	initial	harm	general	initial
erase	general	initial	murder	general	initial
hostile	general	initial	violent	general	initial
violence	general	initial	marginal	general	initial
damage	general	initial	destroy	general	initial
destruction	general	initial	destruct	general	initial
looting	theft	initial	loot	theft	initial
theft	theft	initial	trafficking	theft	initial
illegal	illegal	initial	willful	illegal	initial
black market	illegal	initial	terrorist	conflict	initial
conflict	conflict	initial	shooting	conflict	initial
shoot	conflict	initial	war	conflict	initial
deface	act against property	initial	arson	act against property	initial
graffiti	act against property	initial	inscription	act against property	initial
iconoclasm	act against property	initial	carving	act against property	initial
scratching	act against property	initial	dismantling	act against property	initial
breaking	act against property	initial	break	act against property	initial
abuse elevator	act against property	expanded	scandal	conflict	expanded
genocide ethnic	conflict	expanded	repair amp	act against property	expanded
public property	act against property	expanded	destruction property	act against property	expanded
hate crime	illegal	expanded	infrastructure invade	act against property	expanded
property damage	act against property	expanded	destruction property	act against property	expanded

Table 5: Vandalism lexicon terms.

6. Tweets Detection

In the irrelevant tweets removal section, we have already cleaned datasets by removing irrelevant tweets with predefined irrelevant queries. However, those queries are not exhaustive in capturing

PoS(ISGC2024)009

all forms of irrelevance. Hence, we have eliminated a small portion of irrelevant tweets. Moreover, as our research evolves, a more productive approach would involve determining the criteria for identifying relevant tweets, rather than focusing solely on detecting irrelevance. Irrelevance in disaster-related tweets holds a lot of variability with the introduction of new data samples, while the concept of relevance remains relatively consistent, often linked to a common set of keywords. In this section, we present a task of lexicon-based relevant tweet detection. Based on the developed lexicons, we have defined other tasks to detect tweets, particularly to identify tweets related to natural and human-made disasters, cultural heritage, and vandalism events. Figure 3 shows the number of tweets per category over disaster and cultural heritage domains.

PoS(ISGC2024)009

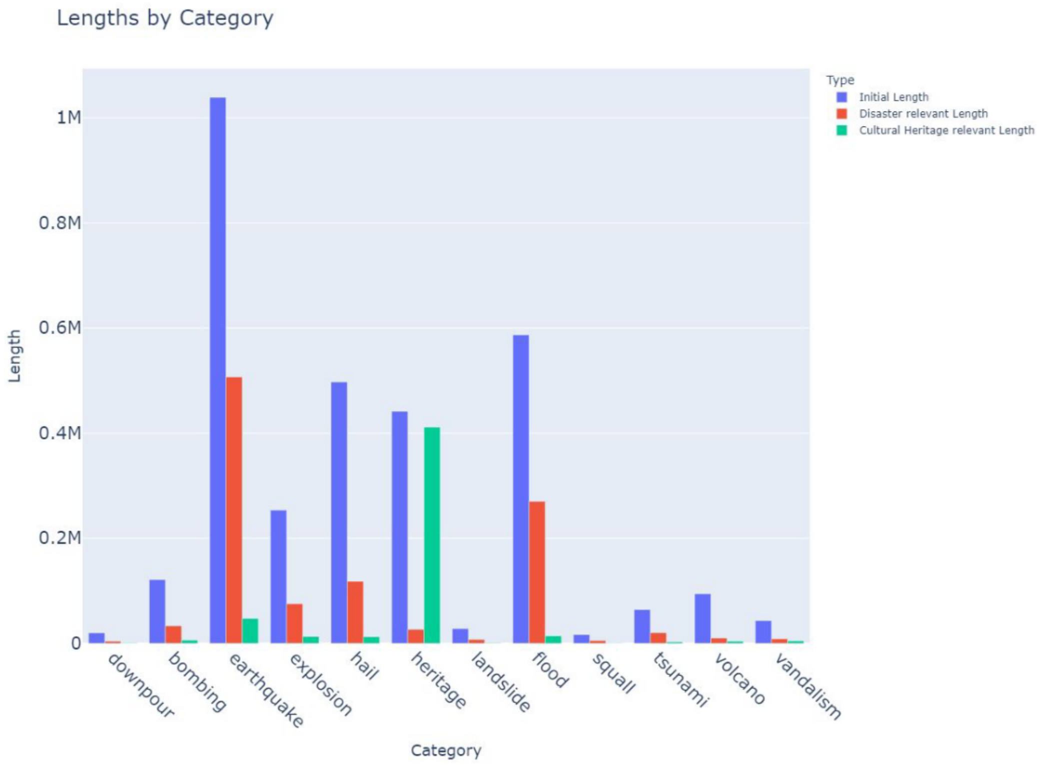

Figure 3: Tweets detection statistics: size by category and detection type.

7. Conclusions

In this study, we have documented the way we have created lexicons for disaster and cultural heritage domains. Due to space constraints, we have not provided our reasoning for the lexicon of the vandalism domain.

In the future, we aim to create a map-based visualization to display the geographic distribution of tweets, allowing us to observe the concentration of Twitter activity in different regions. Furthermore, we intend to integrate sentiment analysis results into the map to provide insights into the emotional tone expressed by users in each geographical area.

References

[1] A. Kruspe, J. Kersten and F. Klan, *Review article: Detection of informative tweets in crisis events*, .

[2] M.L. Jamil, S. Pais and J. Cordeiro, *Detection of dangerous events on social media: A perspective review*, 2204.01351.

[3] P. Kumar, *Twitter, disasters and cultural heritage: A case study of the 2015 nepal earthquake*, *Journal of Contingencies and Crisis Management* **28** (2020) 453.

[4] M. Wang and G. Hu, *A novel method for twitter sentiment analysis based on attentional-graph neural network*, *Information* **11** (2020) 92.

[5] F.A. Lovera, Y.C. Cardinale and M.N. Homsi, *Sentiment analysis in twitter based on knowledge graph and deep learning classification*, *Electronics* **10** (2021) 2739.

[6] I. Temnikova, C. Castillo and S. Vieweg, *Emterms 1.0: A terminological resource for crisis tweets*, in *International Conference on Information Systems for Crisis Response and Management*, 2015, https://api.semanticscholar.org/CorpusID:36168913.

[7] L.A.B. Guerzo, H.A.O. Kilkenny, R.N.D. Osorio, A.H.E. Villegas and C.S. Ponay, *Topic modelling and clustering of disaster-related tweets using bilingual latent dirichlet allocation and incremental clustering algorithm with support vector machines for need assessment*, in *2021 International Conference on Software Engineering amp; Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, IEEE, Aug., 2021, DOI.

[8] W.L. Hamilton, K. Clark, J. Leskovec and D. Jurafsky, *Inducing domain-specific sentiment lexicons from unlabeled corpora*, 1606.02820.

[9] K. Labille, S. Gauch and S. Alfarhood, *Creating domain-specific sentiment lexicons via text mining*, 2017.

[10] J. Qiang, Z. Qian, Y. Li, Y. Yuan and X. Wu, *Short text topic modeling techniques, applications, and performance: A survey*, *IEEE Transactions on Knowledge and Data Engineering* **34** (2022) 1427.

[11] S. Priya, M. Bhanu, S.K. Dandapat, K. Ghosh and J. Chandra, *Taqe: Tweet retrieval-based infrastructure damage assessment during disasters*, *IEEE Transactions on Computational Social Systems* **7** (2020) 389.

[12] A. Olteanu, C. Castillo, F. Diaz and S. Vieweg, *Crisislex: A lexicon for collecting and filtering microblogged communications in crises*, *Proceedings of the International AAAI Conference on Web and Social Media* **8** (2014) 376.

[13] X. Liang, Y. Lu and J. Martin, *A review of the role of social media for the cultural heritage sustainability*, *Sustainability* **13** (2021) 1055.

PoS(ISGC2024)009

[14] C. Sporleder, *Natural language processing for cultural heritage domains*, *Language and Linguistics Compass* **4** (2010) 750.

[15] L.R. Williams, *Vandalism to cultural resources of the rocky mountain west*, .

[16] D. Chatzigiannis, *Vandalism of cultural heritage: Thoughts preceding conservation interventions*, *Change Over Time* **5** (2015) 120.

PoS(ISGC2024)009