A lexicon-based approach to analyse short texts from social media

Summer Student Fellowship Report

Anastasiya Sopyryaeva
anastasiy.sopyryaeva@studio.unibo.it

September 13, 2023

Content

1. Introduction
2. Implementation
    - Domain Lexicons generation
    - Tweets detection
    - Text corpus analysis
3. Further research

Introduction

Introduction

Aim of the activity

To analyze social media texts (e.g., Twitter) to uncover any information valuable for Cultural Heritage Management in Crisis Situations that are triggered by either natural or human-induced disasters (e.g., war, vandalism).

To develop methodologies that leverage AI technologies to efficiently extract this valuable information from the datasets, such as topics of discussion, sentiments expressed by people, or needs expression.

Introduction

Purpose of a lexicon for NLP applications

- *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 and development of AI applications in the research field

Introduction

The objectives defined

1. Create a lexicon of keywords related to the cultural heritage domain
2. Create a lexicon of keywords related to the vandalism domain
3. Adjust open-source crisis lexicons to enable their application to our data

Lexicon	Examples	
	Fundamental vocabulary	**Social media vocabulary**
Cultural Heritage	'manuscript', 'ancient site'	'future generations'
Vandalism	'vandal', 'property damage'	'hate crime', 'sexual abuse', 'ethnic genocide'
Crisis	'injury', 'flood crisis'	

Table 1. Examples of lexicons content

Introduction

The objectives defined

4. Lexicon application to Natural Language Processing tasks
   - Detection of tweets relevant to the domain
   - Topic Modeling

5. Textual features extraction
   - Hashtags
   - Emojis
   - Named Entities

Introduction

Data description

We collected 13 datasets extracted from Twitter API by 13 categories related key terms. The categories chosen are the following.

- 11 categories related to different disaster events: ***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 is composed as follows:

'category value' lang:en -is:retweet -is:reply

The data have been collected from Jan 1 to Apr 26, 2023.

Implementation

Implementation

Data exploration

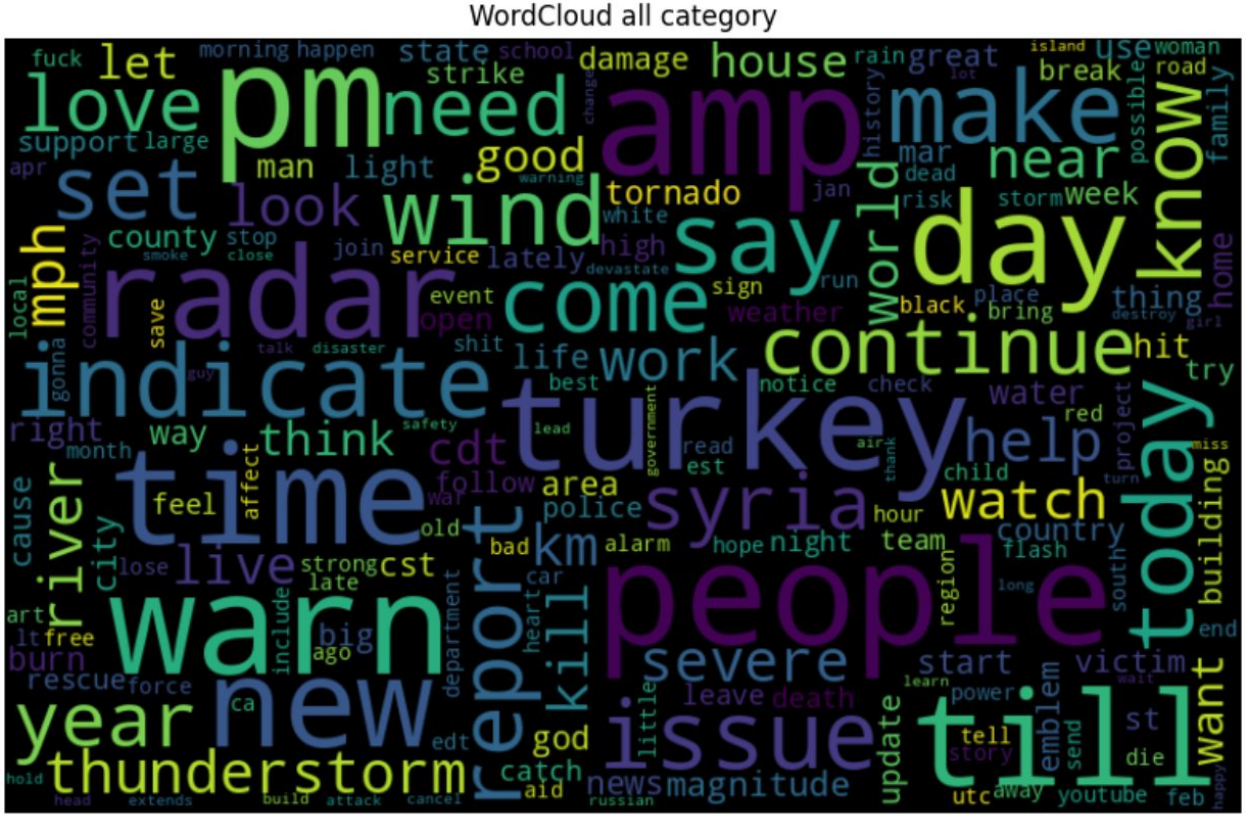


Figure 1. WordCloud Visualisation of unigrams in the text corpus

Implementation

Data exploration

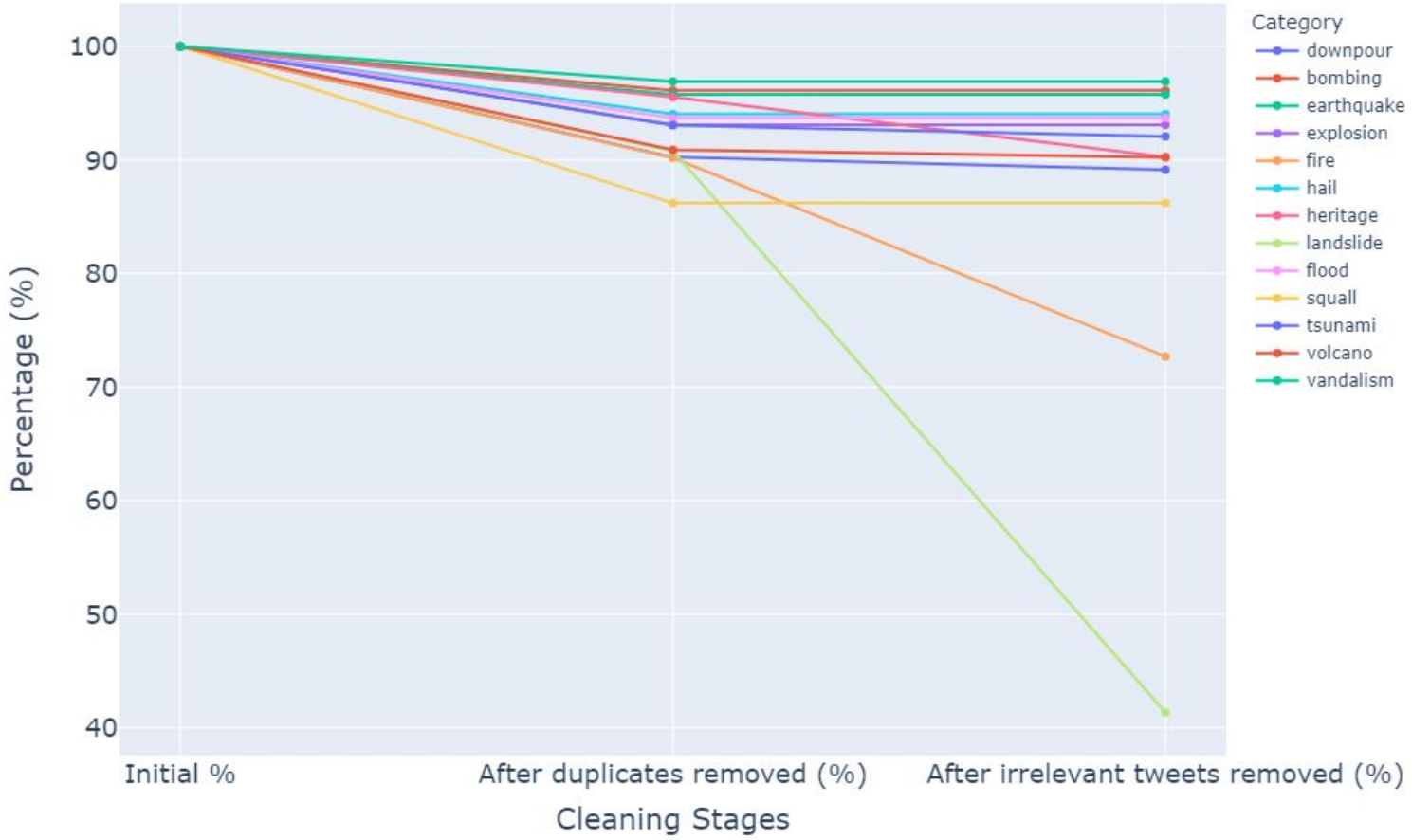
- Duplicates
- Abbreviations
- Complex semantics of terms

Example: “Belarus election: Lukashenko’s claim of **landslide** victory sparks widespread protests”
Example: “This dress is perfect on you! Looking **FIRE**!!”

Category	Irrelevant terms
landslide	fleetwood mac, 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

Table 2. Example of irrelevant vocabulary

Figure 2. Percentage Dataset Size Decrease After Each Cleaning Stage



Emphasizing the lexicon importance

Lexicon Generation

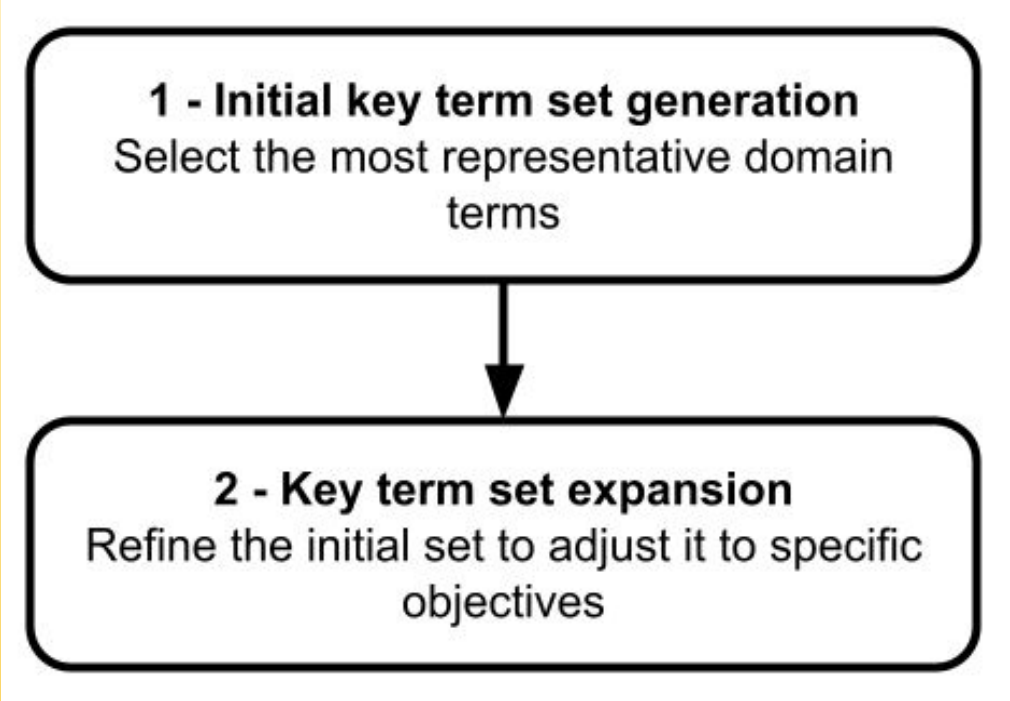


Figure 3. General lexicon generation workflow

Typically, the approach to lexicon generation involves 2 main steps: the initial query generation step, followed by the query expansion step (Olteanu, Castillo, Diaz, Vieweg, 2014).

Adhering to the categorization of our collected datasets into disasters, vandalism, and cultural heritage, we constructed three corresponding lexicons.

Implementation: Crisis Lexicon

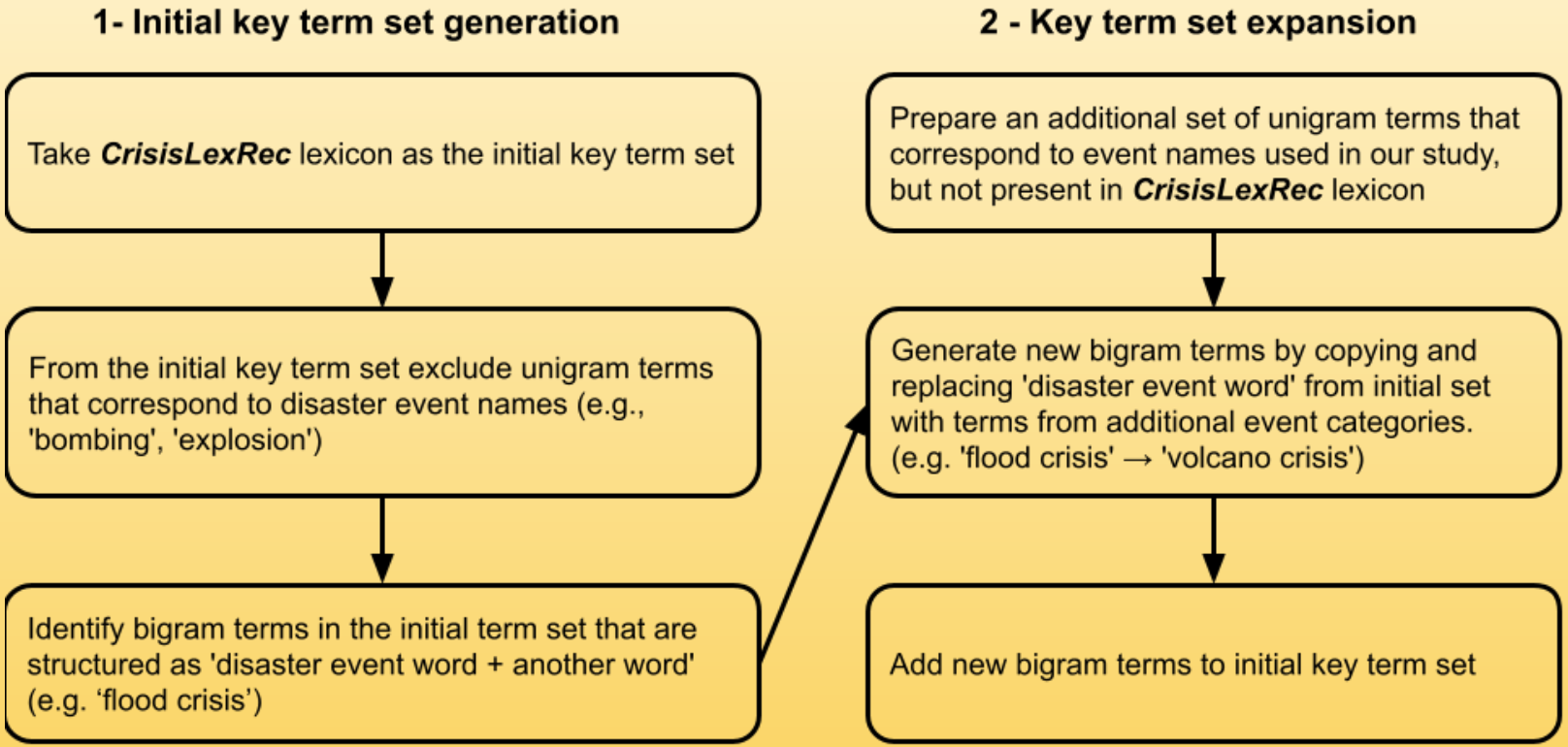


Figure 4. Crisis lexicon generation workflow

Implementation: Crisis Lexicon

CrisisLexRec (Olteanu, et al., 2014) is a crisis lexicon developed upon the CrisisLexT6 collection of tweets (Olteanu, et al., 2014).

CrisisLexT6 tweet collection:

- Includes English tweets across 6 large events in 2012 and 2013
- Contains 60,000 tweets
- About 10,000 tweets labeled by relatedness (as "on-topic", or "off-topic") with each event

CrisisLexRec lexicon:

- 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’, ‘victims’)

Implementation: Crisis Lexicon

Terms	
From CrisisLexRec	**Generated**
flood victim →	earthquake victim
flood powerful →	earthquake powerful
terrified hurricane →	terrified earthquake

Table 3. Example of Crisis Lexicon Terms

Implementation: Cultural Heritage Lexicon

Lexicon generation for cultural heritage (CH) domain →

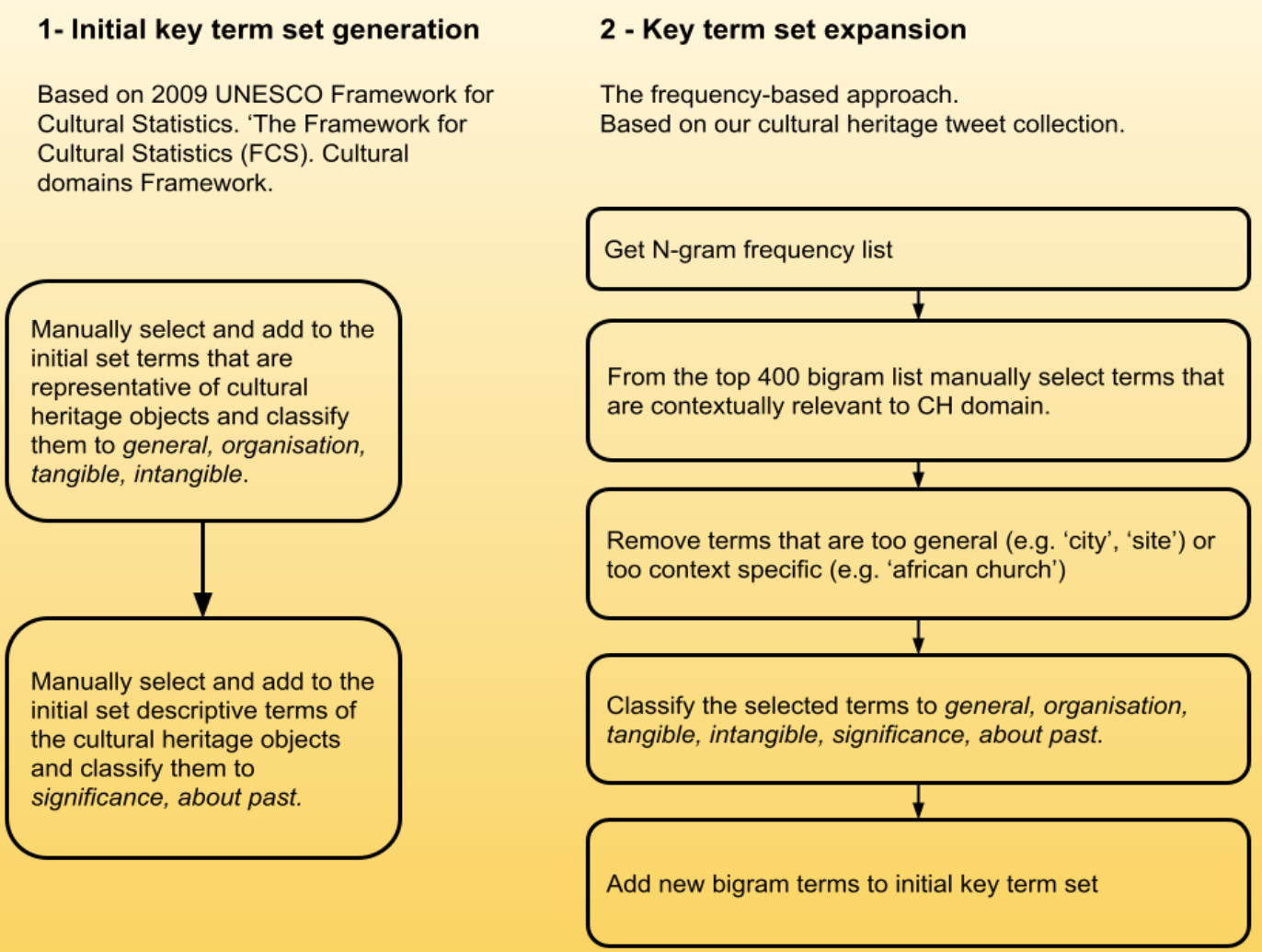


Figure 5. Cultural Heritage lexicon generation workflow

Implementation: Cultural Heritage Lexicon

According to UNESCO Institute for Statistics, “Cultural heritage includes **artefacts, monuments**, a group of **buildings** and **sites**, **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), **intangible** cultural heritage (ICH) embedded into cultural, and natural heritage artefacts, sites or monuments. The definition excludes ICH related to other cultural domains such as festivals, celebration etc. It covers industrial heritage and cave paintings”.

Implementation: Cultural Heritage Lexicon

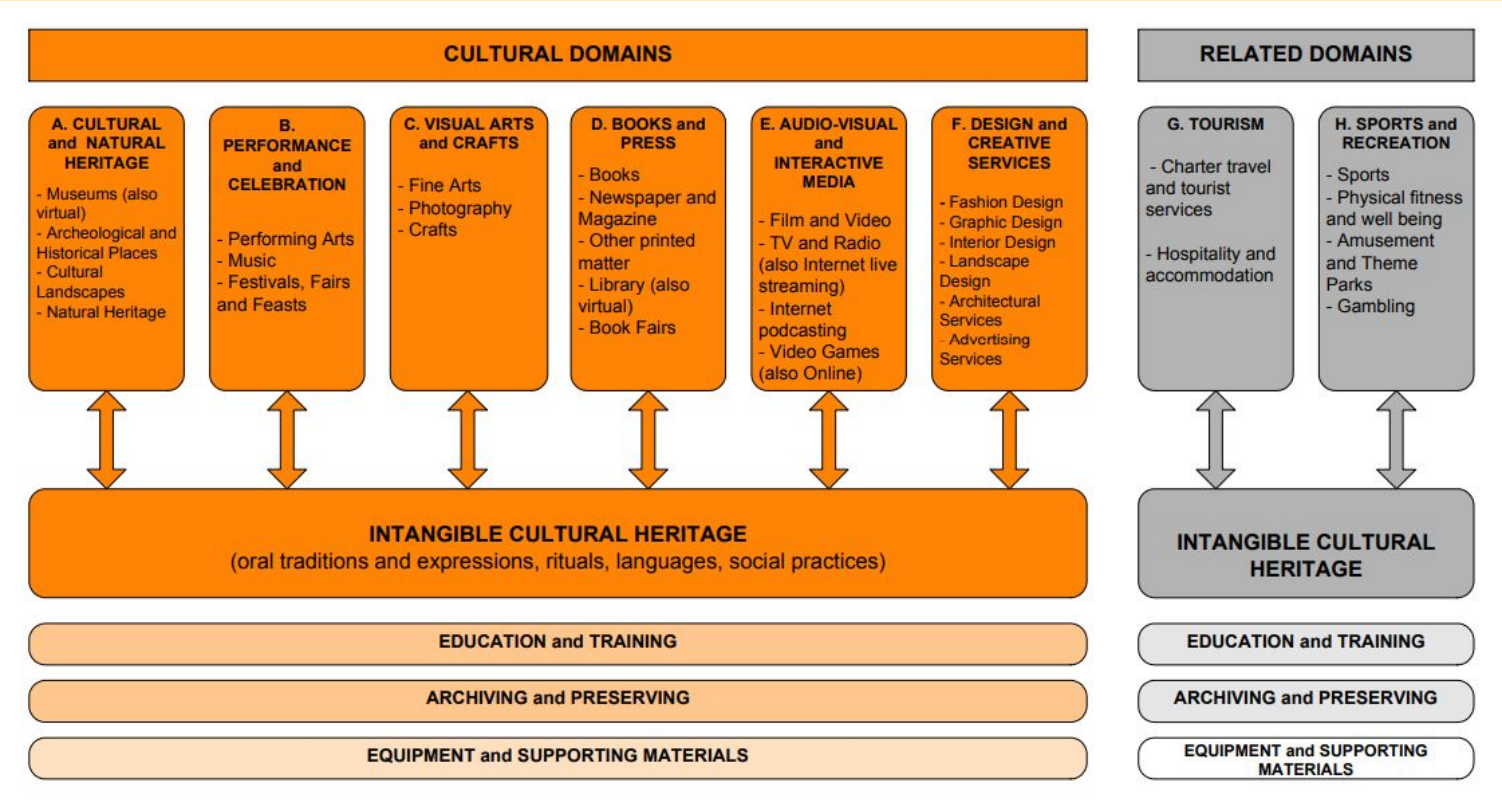


Figure 6.
Framework for cultural statistics domains (from The 2009 UNESCO Framework for cultural statistics (FCS), 2009)

Implementation: Cultural Heritage Lexicon

general: the most generic and definitive terms for cultural heritage domain (e.g., **'heritage'**)
organisation: any organisation/institution that may hold cultural significance (e.g., **'unesco'**)
tangible: terms representing tangible cultural heritage objects (categorised by UNESCO) (e.g., **'musical instrument'**)
intangible: terms representing intangible cultural heritage objects (categorised by UNESCO) (e.g., **'ritual'**)
significance: descriptive terms expressing cultural significance/importance of a phenomenon (e.g., **'symbolic'**)
about past: descriptive terms expressing the historical significance of a phenomenon (e.g., **'ancient'**)

Implementation: Vandalism Lexicon

The lexicon is rooted in the academic literature that investigated vandalism phenomenon from social and cultural perspectives:

Dimitrios Chatzigiannis (Chatzigiannis, 2015) discusses sociopolitical and aesthetic aspects of the vandalism phenomenon from the conservator's point of view. Providing different types, angles and perceptions of vandalism act, the paper contains definitions with diverse vocabulary that is useful for lexicon generation.

In another work (Williams, 1978) authors conducted a sociological survey that resulted in a list with forms of actions that are perceived as vandalism actions against cultural heritage by respondents. These categorisations of actions as understood by people is helpful when adapting lexicon to social media context.

Implementation: Vandalism Lexicon

Lexicon generation for vandalism domain →

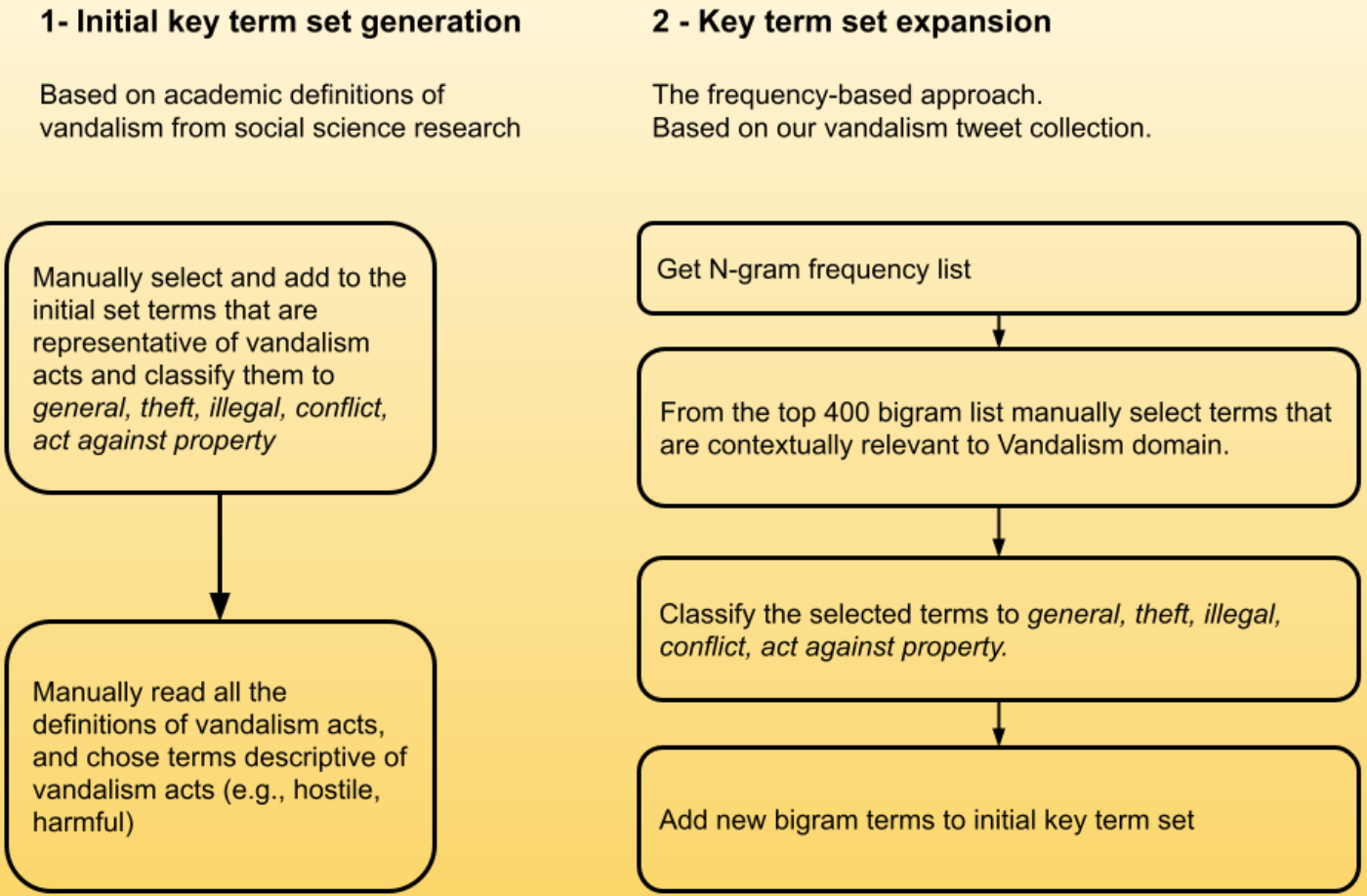


Figure 7. Vandalism lexicon generation workflow

Implementation: Vandalism Lexicon

general: the most generic and definitive 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'***)

Implementation: Tweets detection

Lexicon-based tweet detection

Based on the 3 lexicons we developed in the previous section, now we define the following 3 subtasks related to tweet detection.

Task 1: Among tweets in all datasets identify those that relate to natural and human-made disasters.

Task 2: Among the tweets in datasets related to different types of natural and human-made disasters, identify those that relate to cultural heritage.

Task 3: Among the tweets in all datasets, identify those that relate to vandalism.

Implementation: Tweets detection

Tweet detection workflow →

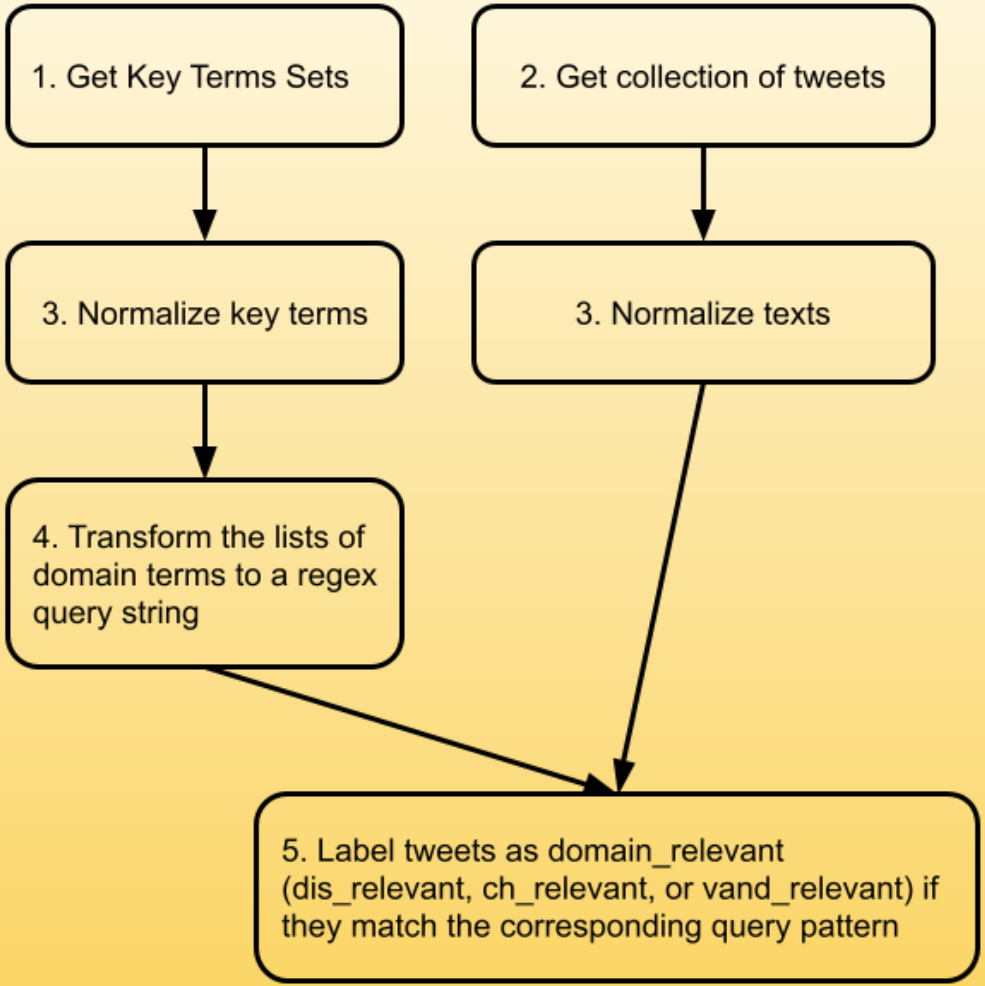
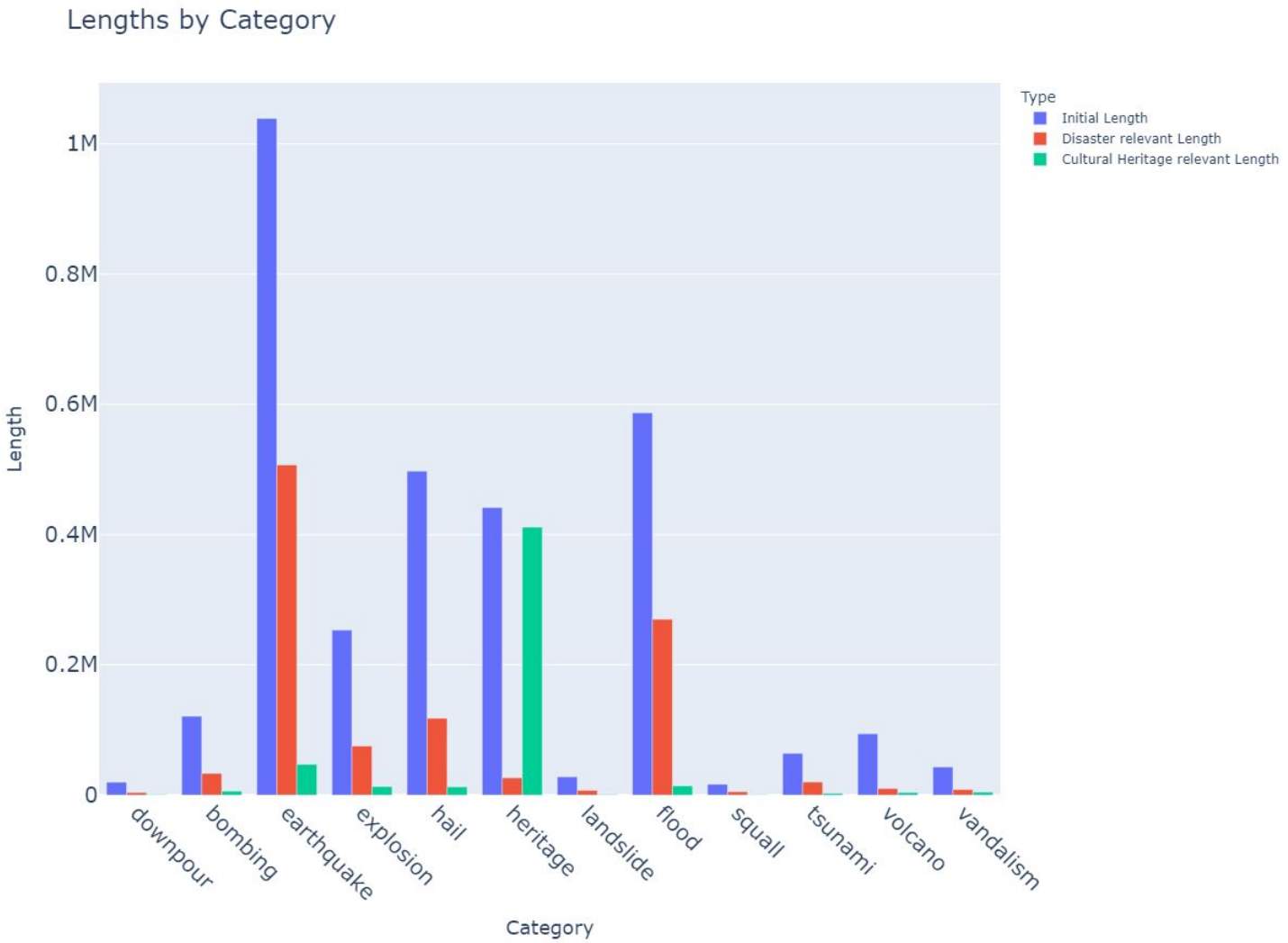


Figure 8. Tweet detection workflow

Figure 9. Tweets detection statistics:

Size by category and detection type



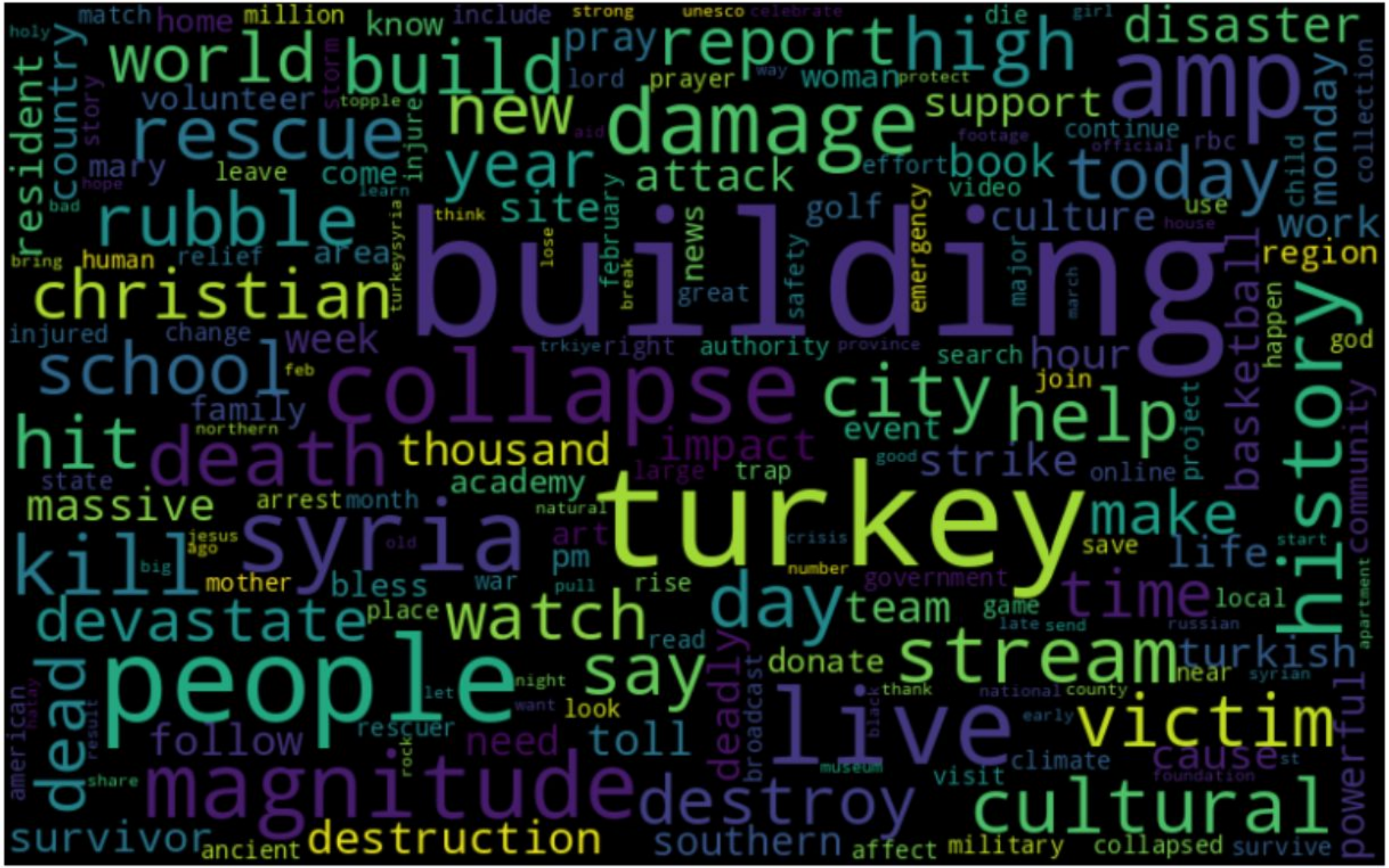


Figure 10. WordCloud Visualisation of unigrams in the text corpus (disaster and cultural heritage relevant)

Implementation: Named Entity Recognition

Experimental setup

The Named Entity Recognition (NER) model first identifies an entity and then categorizes the entity into the most suitable class. (location, person name ect).

- TweebankNLP pre-trained model for NER developed specifically for named entities detection *in tweets* (Jiang, Hua, Beeferman, Roy, 2022).
- The following 4 Named Entities are considered in this model: LOC (location), PER (person), ORG (organisation), MISC (named entities which particular category could not be recognised).
- The model shows 74.35% Entity-Level F1.

Implementation: Named Entity Recognition

Results

The following examples of named entity recognition applied to *'volcano'* category dataset illustrate the potential use of NER for research objectives of this study:

- For location retrieval from texts (LOC)
- For understanding the organizations the Twitter users refer to - sometimes they are irrelevant. For example, recognition of a football club named Volcano point on irrelevance of this tweets to our domain (ORG). Example:

Tweet	**Named Entities**	**Types of the Named Entities**	**Confidence scores**
'The worlds largest volcano **Mauna Loa** is located in **Hawaii** and measures over 30000 feet 9144 meters from its base on the ocean floor'	['Mauna Loa', 'Hawaii']	['LOC', 'LOC']	[0.669804, 0.9257064]
'Lemme move to **Volcano FC**'	['Volcano FC']	['ORG']	[0.8208096]

Table 4. Example of NER output

Implementation: Topic Modeling

Experimental setup

We have selected 2 models to compare and choose the one which performs the best on our datasets.

1. **BERTopic modeling.** A technique based on the state of art transformers model. BERTopic generates document embedding with pre-trained transformer-based language models, clusters these embeddings, and finally, generates topic representations with the class-based TF-IDF procedure (Grootendorst, 2022).
2. **DMM**. A dirichlet multinomial mixture model-based approach. DMM is a probabilistic generative model for documents, and embodies two assumptions about the generative process: (1) the documents are generated by a mixture model, and (2) there is a one-to-one correspondence between mixture components and clusters (Yin, Wang, 2014).

Evaluation method: c_v topic coherence

Implementation: Topic Modeling

Results

category	BERTopic coherence	GMM coherence
bombing	0.8074	0.4518
downpour	0.6198	0.5657
earthquake	0.6593	0.4768
explosion	0.7687	0.4589
flood	0.7574	0.4916
hail	0.8849	0.5616
heritage	0.7860	0.5667
landslide	0.3421	0.5247
tsunami	0.7664	0.4425
volcano	0.4444	0.3569
vandalism	0.7549	0.4347

Table 5. C_V coherence score comparison

Implementation: Topic Modeling

Results

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07
1	history	usaenemyofpeace	rightwing	stophazaragenocide	western	two	oklahoma
2	building	advance	tim	hazaras	prowestern	war	city
3	attack	although	mcveighs	highway	provocation	conscription	terrance
4	people	overtaken	oklahoman	kidnapping	erase	distil	yeakey
5	amp	longer	extremist	increasingly	narrative	letter	allegedly
6	city	aviation	terrify	indiscriminate	ignore	depression	anyone
7	kill	yemen	federal	face	worry	coup	listen
8	war	reason	year	september	medium	revolution	scene
9	us	dresden	building	cultural	foreign	document	commit
10	one	may	raw	religious	hide	interest	officer
Documents count	1684	157	43	27	17	16	15

Table 6. Topics discovered by BERTopic in bombing category tweets (disaster and cultural heritage relevant)

Further research

Further research: Geo Analysis

- Create a map-based visualization to display the geographic distribution of tweets, allowing us to observe the concentration of Twitter activity in different regions
- Integrate sentiment analysis results into the map to provide insights into the emotional tone expressed by users in each geographical area
- Incorporate topic modeling outcomes into the map visualization to highlight the prevalent subjects of discussion in different regions. To provide more details, this can involve:
  - Percentage of each topic
  - Number of messages on each topic
- Integrate other features extracted from the texts, such as:
  - Top N hashtags
  - Top N keyword
  - Most mentioned entities

Further research: Lexicon Refinement and Application

- Improve the lexicons by adding more terms discovered from initial seed set of relevant tweets
- Apply classification techniques on relevant tweets to label them based on topics of discussions and sentiments expressed. The labels are possible to define from the Lexicons and Topic Modeling results

Thank You
for Your Attention!

REFERENCES - 1

[1] A. Olteanu, C. Castillo, F. Diaz, S. Vieweg. CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM'14). AAAI Press, Ann Arbor, MI, USA, 2014.

[2] I. Temnikova, C. Castillo, and S. Vieweg. EMTerms 1.0: A Terminological Resource for Crisis Tweets. In Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM'15). Kristiansand, Norway, 2015.

[3] UNESCO. “UNESCO in brief”. (accessed 10.09.2023). URL: https://www.unesco.org/en/brief

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

[5] UNESCO Institute for Statistics. “Cultural heritage. Definition”. (accessed 10.09.2023). URL: https://uis.unesco.org/en/glossary-term/cultural-heritage

[6] Grammarist. “Landslide”. (accessed 10.09.2023). URL: https://grammarist.com/idiom/landslide/

[7] Slang.net. “Fire”. (accessed 10.09.2023). URL: https://slang.net/meaning/fire

[8] L. R. Williams. Vandalism to Cultural Resources of the Rocky Mountain West. U.S.D.A. Forest Service, Southwestern Region, Cultural Resources Report No. 21, Albuquerque, 1978

REFERENCES - 2

[9] Chatzigiannis, Dimitrios. "Vandalism of Cultural Heritage: Thoughts Preceding Conservation Interventions." Change Over Time 5, no. 1 (2015): 120-135.

[10] Hang Jiang, Yining Hua, Doug Beeferman, Deb Roy. Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis. In Proceedings of the 13th Language Resources and Evaluation Conference (LREC), 2022.

[11] Prateek Majumder. “Named Entity Recognition (NER) in Python with Spacy”.Analytics Vidhya, 2021. (accessed 10.09.2023). URL: https://www.analyticsvidhya.com/blog/2021/06/nlp-application-named-entity-recognition-ner-in-python-with-spacy/

[12] Dat Quoc Nguyen, Thanh Vu, Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020.

[13] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. 2022.

[14] R.J.G.B. Campello, D. Moulavi, J. Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg.

REFERENCES - 3

[15] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. The Journal of Open Source Software, 3(29):861.

[16] Jipeng Qiang et al. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. 2021

[17] J. Yin, J. Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 233–242.

[18] Tong Wei. Terminology and ontology for cultural heritage : application to chinese ceramic vessels. Formal Languages and Automata Theory [cs.FL]. Université Grenoble Alpes [2020-..], 2020.

[19] X. Liang, Y. Lu, J. Martin. A Review of the Role of Social Media for the Cultural Heritage Sustainability. Sustainability 2021, 13, 1055.

[20] Caroline Sporleder. Natural Language Processing for Cultural Heritage Domains. Saarland University. 2010

[21] Pakhee Kumar. Twitter, disasters, and cultural heritage: A case study of the 2015 Nepal earthquake. Journal of Contingencies and Crisis Management, Wiley, 2020.

REFERENCES - 4

[22] Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., and Rana, O.: Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter, Sociological Research Online, 18, 7, 2013

[23] Anna Kruspe, Jens Kersten, and Friederike Klan. Review article: Detection of informative tweets in crisis events,2020.

[24] M. Luqman Jamil, Sebastião Pais, and João Cordeiro. Detection of Dangerous Events on Social Media: A Perspective Review,2022.

[25] Mingda Wang and Guangmin Hu. A Novel Method for Twitter Sentiment Analysis Based on Attentional-Graph Neural Network. Information, 11, pp. 92, 2020.

[26] Fernando Andres Lovera, Yudith Coromoto Cardinale, and Masun Nabhan Homsi. Sentiment Analysis in Twitter Based on Knowledge Graph and Deep Learning Classification. Electronics, 2021.

[27] Lady Angelica Buen Guerzo, Hans Aaron O. Kilkenny, Raphael Noel D. Osorio, Andrei Hart E. Villegas, and Charmaine 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. 2021 International Conference on Software Engineering Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), IEEE, 2021.

REFERENCES - 5

[28] Shalini Priya et al. TAQE: Tweet Retrieval Based Infrastructure Damage Assessment During Disasters. Indian Institute of Technology Patna, India

[29] Hamilton WL, Clark K, Leskovec J, Jurafsky D. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. Proc Conf Empir Methods Nat Lang Process. 2016

[30] Labille, Kevin, Susan Gauch and Sultan Alfarhood. “Creating Domain-Specific Sentiment Lexicons via Text Mining.” (2017).