Coastal Resilience with Social Data Analytics: A Design Science Approach

Carol Lee
Northeastern University, car.lee@northeastern.edu

YoungHo Yoon
University of Massachusetts Boston, youngho.yoon001@umb.edu

Pratyush Bharati
University of Massachusetts, pratyush.bharati@umb.edu

Coastal Resilience with Social Data Analytics: A Design Science Approach

Short Paper

Carol Lee
Northeastern University
360 Huntington Ave, Boston, MA 02115
car.lee@northeastern.edu

Youngho Yoon
University of Massachusetts Boston
100 Morrissey Blvd, Boston, MA 02125
YoungHo.Yoon001@umb.edu

Pratyush Bharati
University of Massachusetts Boston
100 Morrissey Blvd, Boston, MA 02125
Praytush.Bharati@umb.edu

Abstract

We adapt a design science approach (DSR) for coastal resilience and climate justice using big data analytics. Our big data and machine learning based artifact can accommodate various sets of social attributes to understand coastal risks for vulnerable communities. We analyzed social data from communities vulnerable to coastal hazards by incorporating machine learning (ML) to assess coastal community needs and demands. In addition, we developed a user interface that provides data selection and weighting functionalities. We extend IS literature in design science research and ML techniques to further our understanding of coastal resilience in vulnerable communities. The outcomes of this research can help community members and policy makers understand and develop robust sustainability and climate focused decisions using a coastal resilience decision approach.

Keywords: Coastal resilience, big data analytics, design research approach, machine learning, climate justice, social data

Introduction

We apply design science to address coastal resilience with a focus on climate justice using big data analytics and machine learning (ML). We adopt a design science research (DSR) approach (Hevner et al., 2004; Peffers et al., 2007) to develop a coastal resilience artifact that analyzes climate resilience data sets, including social data from communities vulnerable to coastal hazards in the greater Boston area in the U.S. This research aims to provide a more comprehensive coastal resilience decision making system that integrates the coastal communities' voices. Globally, a billion people are projected to be at risk from coastal-specific climate hazards in low-lying cities (IPCC, 2022). Furthermore, these vulnerable coastal communities have historically been disproportionately impacted because of climate change and natural disasters.

More specifically, we aim to identify the key stakeholders, needs and actions of communities and develop a coastal resilience engine for community members and policy makers. We incorporate ML into a flexible analytical approach to assess coastal community needs and demands from social data. Decision makers, who take social and regulatory factors into consideration, can make enhanced climate vulnerability assessments with ML-enhanced models. For instance, social media and sensor data have been analyzed with supervised learning and natural language processing (NLP) to help forecast storms and disaster risks (Harvey et al., 2019; Kankanamge et al., 2020).

We plan to address these challenges by extending information systems (IS) literature in design research and ML techniques to further our understanding of coastal resilience in vulnerable communities. Previous studies in information systems literature focus on warning and disaster systems and do not focus on community decision making (Harvey et al., 2019; Jakariya et al., 2020; Saravi et al. 2019). Few research studies have explored systems that integrate social data related to climate change or disaster resilience (Enenkel et al., 2020; Kankanamge et al., 2020; Lee et al., 2022; Seidel et al., 2018). Our research approach, enhanced with social data, can accommodate various sets of social attributes to help us understand coastal risks in vulnerable communities. We propose an innovative artifact that methodically extracts both long-term and short-term objectives, aimed at addressing coastal resilience, directly from the voices in the community. The outcomes of this research can help community members and policy makers understand and develop robust sustainability and climate focused decisions using a coastal resilience decision approach.

Brief Literature Review

We build upon social analytics and decision making IS literature as well as adapt the design science approach. Organizations have employed big data analytics for knowledge-based decisions (Bharati & Chaudhury, 2019) and to understand the influence of social media data on organizational outcomes (Saraf et al., 2022). Although data analytics techniques have been adapted for decision models, incorporating unstructured data, such as images and videos, to address theory and practice is a challenge (Bharati, 2017, pp. 273-276). Social factors play an important role in decision making on sustainability (Carberry et al., 2019) and the role of IS is integral to solving climate change challenges (Seidel et al., 2017). In addition, ML and IS literature can facilitate environmental governance and improve organizational processes and individual practices (Enenkel et al., 2020; Nishant et al., 2020). For example, classification methods were applied to historical flood behaviors to understand flood resilience (Saraviet et al., 2019) and recent research has investigated the usage of social data for coastal resilience (Lee et al., 2022). We extend these decision models to include a process that integrates and analyzes social data focused on coastal resilience.

Research Design

This research adapts a DSR approach for designing and evaluating an information technology (IT) artifact, broadly defined as "constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems)" (Hevner et al., 2004). In alignment with the DSR methodology (Peffers et al., 2007), our approach is designed to be flexible with an outcome-oriented interface for coastal resilience decision-making. The research design includes the development of a coastal resilience approach that employs ML for analyzing social data from coastal communities (see Figure 1).

The process includes the identification of relevant social data for coastal resilience. Then, a multi-step data wrangling and analysis process extracts meaningful keywords that are semantically associated with each other. We adapt Non-Negative Matrix Factorization (NMF), one of the prominent topic modeling techniques introduced by Lee and Seung (1999), to extract human-interpretable topics from the text that are eventually clustered (Casalino et al., 2018). NMF is an important unsupervised learning technique for dimensional reduction and data representation. It is known as the prominent dimensionality reduction technique for text data which returns low-rank factor matrices that uncover the underlying structures and patterns in the data (Lee & Seung, 1999). Specifically, NMF factorizes the original high-dimensional text data matrix into two low-rank matrices that represent the intensity of terms per topic (Xu et al., 2003). NMF is known to produce improved outcomes for sparse and short text data like microblogging sites due to its ability to extract the compact low-rank feature matrix that uncovers the most important features for each topic (Yan et al., 2013). This technique is suitable for our study which intends to extract the underlying key topics from a large data corpus. Lastly, we develop a coastal resilience engine for users to select weights and explore the results.

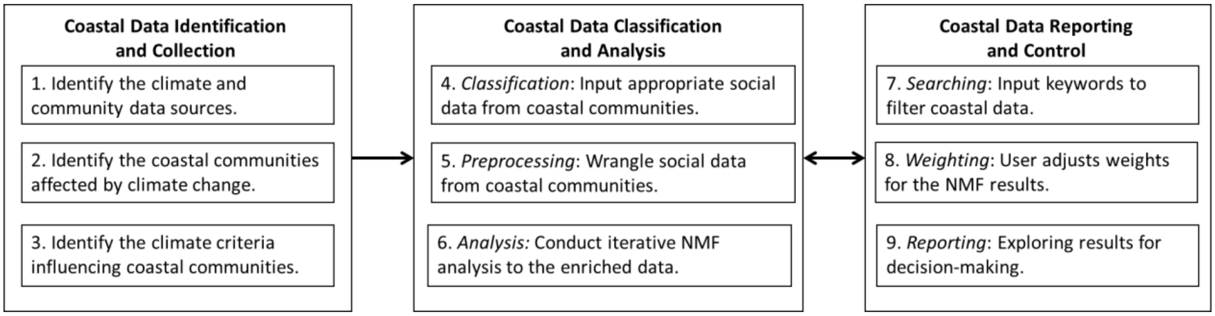


Figure 1. Research Approach

Methodology

Design Requirements

Our IS design fulfills the following design requirements (see Table 1) discussed in previous studies (Zampou et al., 2021). First, the artifact should be capable of capturing and storing real-time social data. Due to the nature of social data, where discussions are dynamically occurring and changing, it is important to collect the most recent data to derive relevant results for topic modeling. To meet the data supply needs for the prototype, we collected social data related to climate change and coastal resilience discussions. Future data collection is possible on a regular basis using the model, and the collected data can be further integrated into the existing data repository.

Second, the collected data needs to undergo systematic classification using a topic modeling technique. Although the data is limited to specific regions and timeframes, the sheer volume of social media records makes manual classification challenging. Therefore, we initially perform data classification based on the relevance of the content using NMF topic modeling. This approach can facilitate easier data selection and representation in the future.

Last, the reporting and control of the classified content should be user-friendly in the front-end layer, which is presented to end users. The classified data should be provided in a format that users can comprehend, and a certain level of controllability is required in the result generation process and interface. In this interface, we designed three screens with tables containing example tweets and metrics that can help the user understand topics from the social data. We also have included a control panel that allows users to directly adjust the weights of features used for data selection, enabling them to receive more satisfactory results in the reported outcomes.

Design Requirement	Description
Data Identification and Collection	Collecting social data related to coastal resilience, where identifying relevant social data in real-time is important for decision-making.
Data Classification and Analysis	Processing and analyzing the sheer volume of social media data into several clusters or groups based on the similarity of semantic structures found in the relationships between key terms.
Data Reporting and Control	Reporting key findings, topics, and tweets, filtered by the topic modeling algorithm and topic selection made by users, in user-friendly interpretable and controllable ways.

Table 1. Design Requirements

Data Identification and Collection

For this research, we collected data related to discussions on coastal resilience and climate justice within the community and local environmental organizations. We identified coastal-related accounts, including non-governmental organizations (NGOs), social movement organizations (SMOs), as well as individual users, all of whom were associated with the greater Boston area. The data collection process involved utilizing coastal-related keywords and the names of environmental groups and coastal communities to gather relevant data from their respective accounts. Our preliminary dataset consists of 18,857 tweets that were generated in the year 2021. Each tweet in the dataset includes various features such as the full text, timestamp, tweet metrics (e.g., number of likes, retweets), user metrics (e.g., number of followers), and URL, among others. To ensure user data privacy, our interface does not display identifiable information (e.g., usernames), and we have emphasized the importance of not sharing or reposting this data to mitigate any potential privacy or ethical concerns.

Data Classification and Analysis

Before conducting the analysis, we preprocessed the text to ensure it was consistent for NLP. Text preprocessing is a crucial step in NLP that involves a series of techniques to clean, tokenize, normalize, and transform the text data (Uysal et al., 2014). Our preprocessing steps included tokenization to break the text into smaller units, lowercasing, removing unnecessary characters (such as symbols, punctuation, and URLs), and stemming and lemmatization to extract nouns. We also removed stopwords (e.g., "and", "the", "is") and high-frequency words (e.g., “Massachusetts”, “Boston”, “election”, “journalism”) and media names that could potentially mislead from coastal change. We also computed a word frequency list for the entire data set for robustness. For example, "action" appeared 2,775 times, "city" appeared 2,189 times, and "crisis" appeared 1,903 times.

We employed NMF, introduced by Lee and Seung (1999), for conducting a topic modeling that extracts human-interpretable topics from the social data using the NMF module in the scikit-learn package (Cichocki & Phan, 2009) and NLTK package. We assigned an ID and text to each document to identify the semantic relationship between words and extracted the key topics for all the documents. The framework of this study is based on a three-stage process (Casalino et al., 2018). The first stage is devoted to dataset creation that transforms a collection of raw data into a preprocessed dataset in a matrix V (Visible Variables). Documents and terms are represented in columns and rows respectively, where values are assigned according to the weighting function of term frequency and inverse document frequency.

In the second stage, our NMF model automatically factorized the data matrix to extract human-interpretable topics by reducing the dimensionality of the dataset into two low-rank factor matrices, which are the document-topic matrix W (weights) and topic-term matrix H (hidden variables). In the W matrix, each row represents a document composed of unnormalized probabilities of topics and each column represents a semantic feature, consisting of a topic. Each value in the W matrix represents the aggregated weights of topic-related terms by each document. In the H matrix, each row represents the distribution of term frequencies in each topic, and each column represents a visible variable, that is, terms. Each topic, driven by the H matrix, provides a set of significant terms that are derived from a collection of bilateral terms that frequently appear together in the same document. As an illustration, one of the topics extracted from our NMF analysis reveals a cluster of semantically related keywords: “sea-level-rise-harbor-island-stormwater-town”.

For each result set, we developed composite scores, which are calculated by combining the topic score (i.e., topic coherence score and occurrence of the search term) and the user impact score (i.e., number of retweets, likes, and followers). Each component score undergoes normalization and logarithmic transformations to eliminate any data skewness or dominance. By default, the topic coherence score is weighted at 70%, and the user impact score is at 30%.

Data Reporting and Control

We designed an IS interface specifically to present relevant tweet data to stakeholders in coastal resilience-affected communities (see Figure 2). We adapted the design principles to follow outcome-oriented principles (Gregor et al., 2020; Seidel et al., 2018). The interface follows three stages, which involve: (1) inputting one or multiple keywords to filter topics from NMF results, (2) selecting topics to load corresponding tweets and adjusting weights for further filtering, and (3) exploring the results and tweets. Each stage is presented on a single screen, providing users with the ability to explore information and control the entire process. To construct the user interface, we utilized Kivy for the interface and Python libraries for text preprocessing.

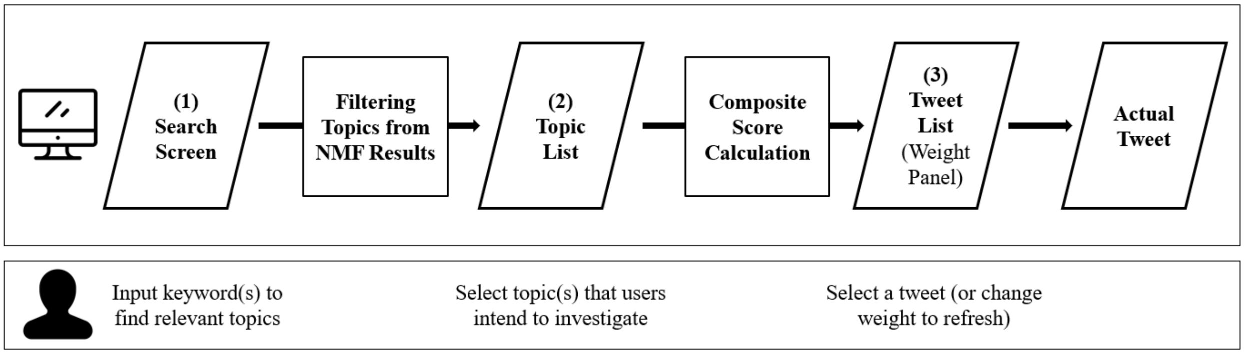


Figure 2. User Interaction with Coastal Resilience Engine

Preliminary Results

In this section, we present both the initial interface and preliminary results of the NMF analysis. Our interface is designed to provide users with relevant data on specific climate issues. The first stage utilizes a search method where users input one or multiple words to explore topics of interest (see Figure 3). Upon clicking the enter button, a list of topics will be displayed, sorted by the frequency of the search words occurring in tweets for each topic.

Alongside each topic, example tweets and percentage scores of word frequency are provided, allowing users to select one or multiple topics for investigation. For instance, using the keyword 'flooding' as an example, the interface generates multiple topics ranked by keyword frequency, indicating the occurrence of the word 'flooding' within a group of tweets for each topic. Our findings reveal that topic numbers 28 and 15 have 199 and 160 keyword-relevant tweets, respectively, with a frequency rate exceeding 10% (see Figure 4).

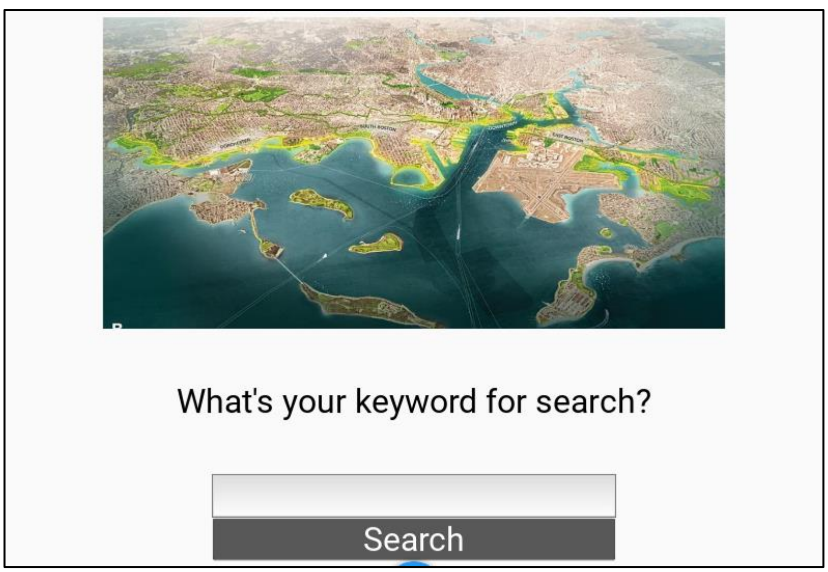


Figure 3. Search Screen in the Interface

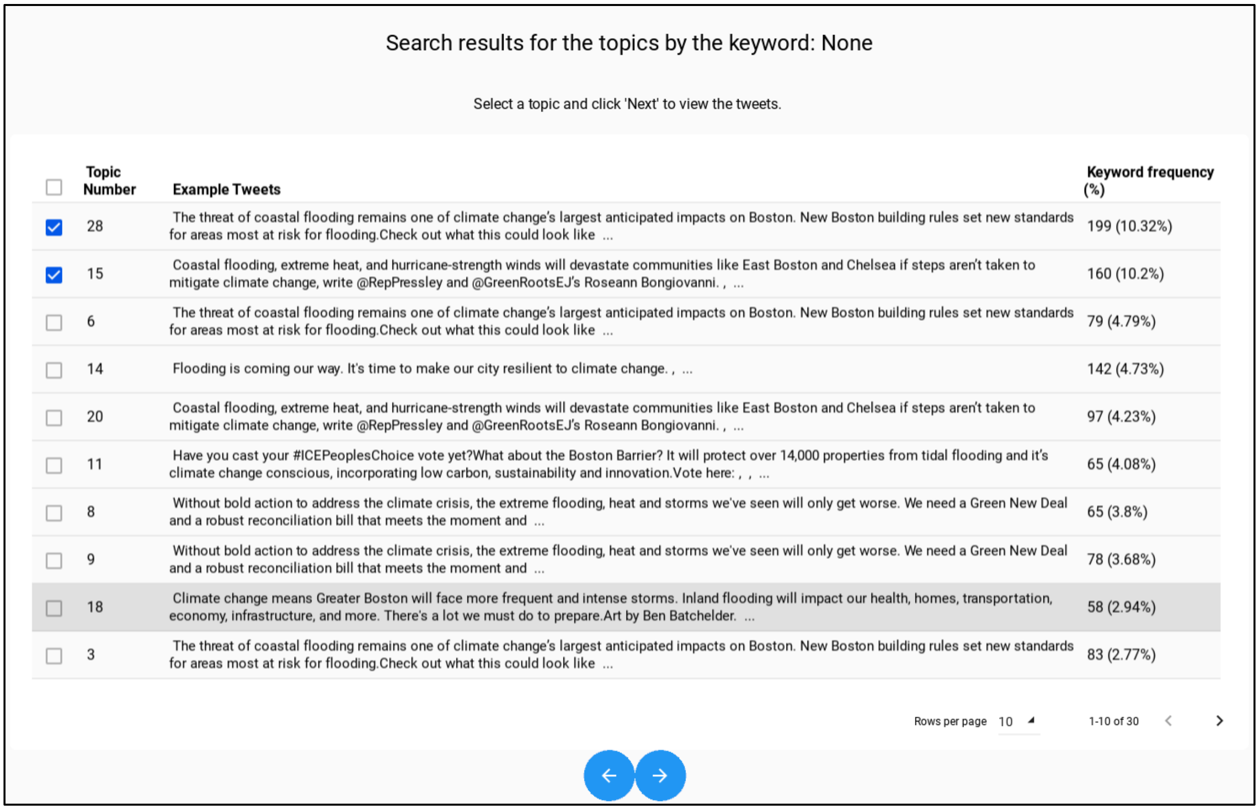


Figure 4. Topic List Screen in the Interface

Based on the sorted results, users can select one or multiple topics based on their preferences. Clicking the "Next" button transitions the screen to the third stage, where users can view a list of tweets sorted by composite scores. Additionally, users have an optional panel to adjust the weight scores for each component of the composite score to improve results as they want. Users can increase either topic coherence scores or user impact scores based on their preference (see Figure 5). For example, users can set the weights of the topic coherence score and search term occurrence to 35% each and the three user impact weights to 10% respectively to see the tweets with higher topic scores. This allows for a personalized user experience and ensures that the interface caters to individual preferences and priorities.

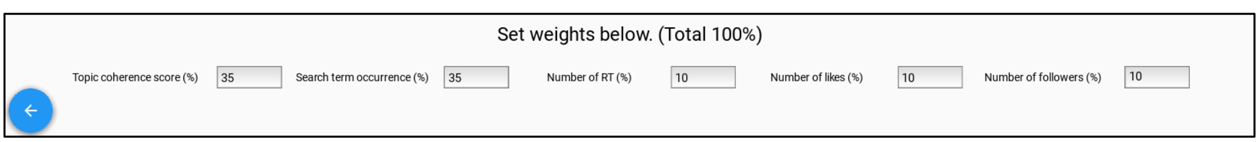
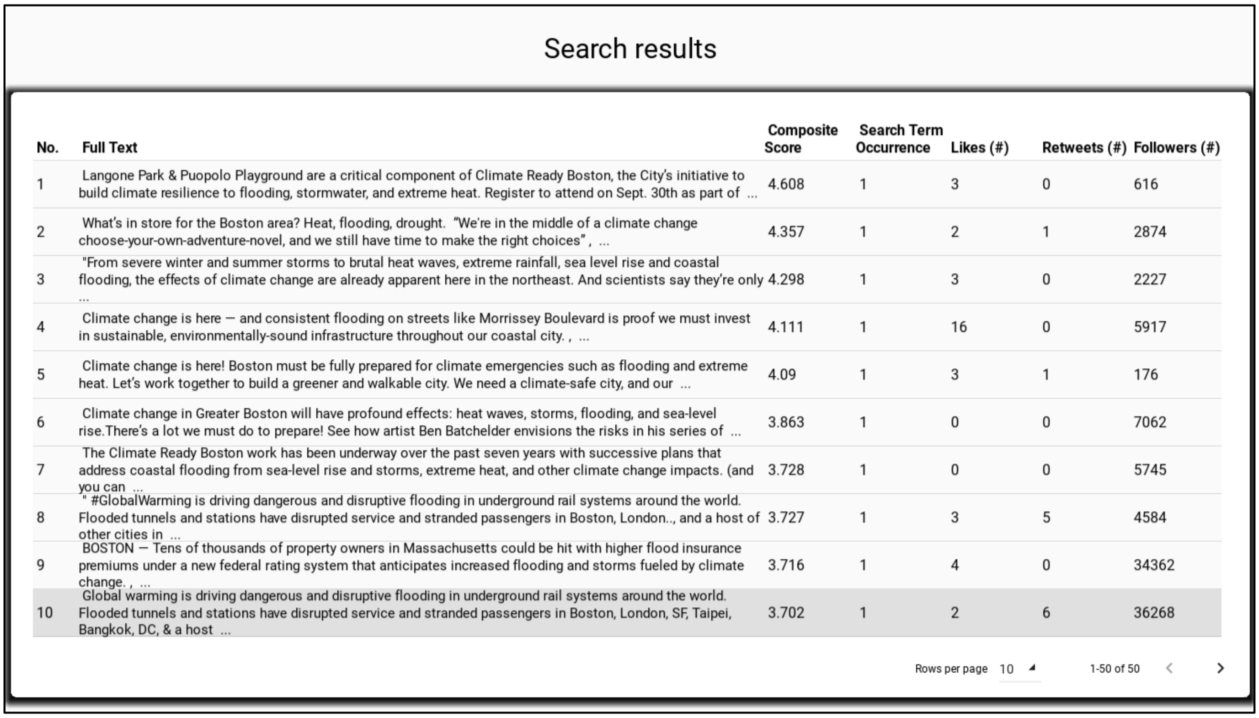


Figure 5. Weight Panel in the Interface

The results include the full text of each tweet, its composite score, and the tweet metrics used for calculating the composite score. Users can access the original tweet by clicking on a specific row (see Figure 6). The outcomes can be modified by adjusting the weight scores through the weight panel. By increasing the topic coherence score and search term occurrence score, while decreasing user impact scores, the ranking of the results undergoes a noticeable transformation. For instance, area-based tweets take precedence with updated rankings, which emphasize vulnerable locations to coastal flooding, such as Chelsea and East Boston.

In the case of our example keyword "flooding," the results display a collection of tweets reflecting the discussions within the Boston community regarding various aspects related to flooding. In our results, tweet number 1 describes the city's initiative in Langone Park located in North End to enhance climate resilience. Tweet numbers 2 and 3 highlight flooding as an indicator of climate change, and tweet number 4 reports specific areas prone to flooding, such as Morrissey Boulevard near the coast. Tweets discussing the impact of flooding on transportation and property insurance are also discussed in the additional high rank tweets.



Search results

No.	Full Text	Composite Score	Search Term Occurrence	Likes (#)	Retweets (#)	Followers (#)
1	Langone Park & Puopolo Playground are a critical component of Climate Ready Boston, the City's initiative to build climate resilience to flooding, stormwater, and extreme heat. Register to attend on Sept. 30th as part of ...	4.608	1	3	0	616
2	What's in store for the Boston area? Heat, flooding, drought. "We're in the middle of a climate change choose-your-own-adventure-novel, and we still have time to make the right choices" , ...	4.357	1	2	1	2874
3	"From severe winter and summer storms to brutal heat waves, extreme rainfall, sea level rise and coastal flooding, the effects of climate change are already apparent here in the northeast. And scientists say they're only ...	4.298	1	3	0	2227
4	Climate change is here — and consistent flooding on streets like Morrissey Boulevard is proof we must invest in sustainable, environmentally-sound infrastructure throughout our coastal city. , ...	4.111	1	16	0	5917
5	Climate change is here! Boston must be fully prepared for climate emergencies such as flooding and extreme heat. Let's work together to build a greener and walkable city. We need a climate-safe city, and our ...	4.09	1	3	1	176
6	Climate change in Greater Boston will have profound effects: heat waves, storms, flooding, and sea-level rise.There's a lot we must do to prepare! See how artist Ben Batchelder envisions the risks in his series of ...	3.863	1	0	0	7062
7	The Climate Ready Boston work has been underway over the past seven years with successive plans that address coastal flooding from sea-level rise and storms, extreme heat, and other climate change impacts. (and you can ...	3.728	1	0	0	5745
8	" #GlobalWarming is driving dangerous and disruptive flooding in underground rail systems around the world. Flooded tunnels and stations have disrupted service and stranded passengers in Boston, London.., and a host of other cities in ...	3.727	1	3	5	4584
9	BOSTON — Tens of thousands of property owners in Massachusetts could be hit with higher flood insurance premiums under a new federal rating system that anticipates increased flooding and storms fueled by climate change. , ...	3.716	1	4	0	34362
10	Global warming is driving dangerous and disruptive flooding in underground rail systems around the world. Flooded tunnels and stations have disrupted service and stranded passengers in Boston, London, SF, Taipei, Bangkok, DC, & a host ...	3.702	1	2	6	36268

Rows per page 10 1-50 of 50

Figure 6. Tweet List in the Interface

Discussion and Future Research

Our preliminary results indicate the needs of communities that are vulnerable to coastal flooding, which can impact the allocation of resources. The ranking of the tweet list in the interface can be used to understand the key topics of coastal communities and engage participants in the decision-making process. Our interface may be used by community members, local organizations, and decision-makers to view and better understand the topics and tweets that concern them and their community.

There are several limitations of our research. First, although we adapt selection rules for the number of topics, we acknowledge all the community voices in our exploratory dataset may not be sufficiently captured. Second, our artifact focuses on providing relevant topics and tweets based on the keyword search, so the context of the specific insights will vary. Thus, future studies will address such issues by integrating complementary sources of community data (e.g., public hearings, workshops, and reports) to encompass a wider spectrum of discussions on coastal resilience and facilitate the development of actionable solutions. Furthermore, we plan to enhance the user interface of the artifact to include user feedback features and further customize the search results based on user preferences.

In this paper, we have presented the analysis results using the design of the artifact and real-world sample data in the initial stages. Our proposed approach is unique in that our artifact methodically extracts both long-term and short-term objectives (i.e., topics) directly from community stakeholders to address coastal resilience. For future research, we intend to conduct a thorough evaluation through interviews and surveys with stakeholders, including environmental experts, scholars, and activists for potential improvements of the artifact. Furthermore, we plan to analyze a larger volume of additional real-world social data integrated with climate observations for developing a coastal decision model leveraging additional techniques, such as developing a neural network to further classify the topics. We will explore the topics in further detail to develop climate justice factors and map them to community data, such as neighborhood, population, and vulnerabilities. We aim to develop additional views and reporting functionalities for decision makers to input coastal change indicators, community information, and policy plans for a multi-view model to address coastal resilience in vulnerable communities. This research can help community members and policymakers make better social data based robust sustainability decisions.

References

Bharati, P. (2017). Big Data Is Not a Monolith. In W. Shen (Eds.), *Book Reviews INFORMS Journal on Applied Analytics (*pp. 273-276), The MIT Press.

Bharati, P., & Chaudhury, A. (2019). Assimilation of big data innovation: Investigating the roles of IT, social media, and relational capital. *Information Systems Frontiers*, *21*, 1357-1368.

Carberry, E. J., Bharati, P., Levy, D. L., & Chaudhury, A. (2019). Social movements as catalysts for corporate social innovation: Environmental activism and the adoption of green information systems. *Business & Society*, *58(5),* 1083-1127.

Casalino, G., Castiello, C., Del Buono, N., & Mencar, C. (2018). A framework for intelligent Twitter data analysis with non-negative matrix factorization. *International Journal of Web Information Systems*, *14(3),* 334-356.

Cichocki, A., & Phan, A. H. (2009). Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer* sciences, 92(3), 708-721.Cichocki, A., and Phan, A. H. 2009. “Fast local algorithms for large scale nonnegative matrix and tensor factorizations,” *IEICE transactions on fundamentals of electronics, communications and computer sciences* (92:3), pp. 708-721.

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. International Journal of Information Management, 60, 102383.Collins, C., Dennehy, D., Conboy, K., and Mikalef, P. 2021. “Artificial Intelligence in Information Systems Research: A Systematic Literature Review and Research Agenda”, *International Journal of Information Management* (60), 102383.

Enenkel, M., Brown, M. E., Vogt, J. V., McCarty, J. L., Bell, A. R., Guha-Sapir, D., Dorigo, W., Vasilaky, K., Svoboda, M., Bonifacio, R., Anderson, M., Funk, C., Osgood, D., Hain, C., & Vinck, P. (2020). Why predict climate hazards if we need to understand impacts? Putting humans back into the drought equation. *Climatic Change, 162*, 1161-1176.

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: the anatomy of a design principle. *Association for Information Systems, 21(6), 1622-1652*.

Harvey, J., Kumar, S., & Bao, S. (2019). Machine Learning-Based Models for Assessing Impacts Before, During and After Hurricane Florence. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 714-721.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28(1),* 75-105.

IPCC. (2022) *Climate Change 2022: Impacts, Adaptation, and Vulnerability*. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press.

Jakariya, M., Alam, M. S., Rahman, M. A., Ahmed, S., Elahi, M. M. L., Khan, A. M. S., Saad, S., Tamim, H. M., Ishtiak, T., Sayem, S. M., Ali, M. S., & Akter, D. (2020). Assessing Climate-induced Agricultural Vulnerable Coastal Communities of Bangladesh using Machine Learning Techniques. *Science of the Total Environment, 742*, 140255.

Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, M. (2020). Determining Disaster Severity Through Social Media Analysis: Testing the Methodology with South East Queensland Flood Tweets. *International Journal of Disaster Risk Reduction*, *42*, 101360.

Lee, C., Yoon, Y., & Bharati, P. (2022). Coastal Resilience Decision Making with Machine Learning. In *Proceedings of Americas Conference on Information Systems 2022*, 3.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401(6755),* 788-791. 102104.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems, 24(3),* 45-77.

Saraf, N., Bharati, P., & Ravichandran, T. (2022). Does Social Capital Arise from Enterprise or Public Social Media Use? A Model of Social Media Antecedents and Consequences. *Information Systems Frontiers*, 1-23.

Saravi, S., Kalawsky, R., Joannou, D., Rivas Casado, M., Fu, G., & Meng, F. (2019). Use of artificial intelligence to improve resilience and preparedness against adverse flood events. *Water, 11(5),* 973.

Seidel, S., Bharati, P., Fridgen, G., Watson, R. T., Albizri, A., Boudreau, M. C. M., Butler, T., Kruse, L.C., Guzman, I., Karsten, H., Lee, H., Melville, N., Rush, D., Toland, J., & Watts, S. (2017). The Sustainability Imperative in Information Systems Research. *Communications of the Association for Information Systems, 40*, pp.1, 3.

Seidel, S., Kruse, L. C., Székely, N., Gau, M., & Stieger, D. (2018). Design principles for sensemaking support systems in environmental sustainability transformations. *European Journal of Information Systems, 27(2),* 221–247.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information processing & management, 50(1),* 104-112.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization, In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval,* 267-273.

Yan, X., Guo, J., Liu, S., Cheng, X., & Wang, Y. (2013). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, 749-757.

Zampou, E., Mourtos, I., Pramatari, K., & Seidel, S. (2022). A design theory for energy and carbon management systems in the supply chain. *Journal of the association for information systems*, *23*(1), 329-371.