A NOVEL NATURAL LANGUAGE PROCESSING APPROACH TO SUPPORT DECISION MAKING FOR ADAPTING CRITICAL INFRASTRUCTURE TO CLIMATE CHANGE

Tanwi Mallick
Assistant Computer Scientist
Mathematics & Computer Science Division
Argonne National Laboratory

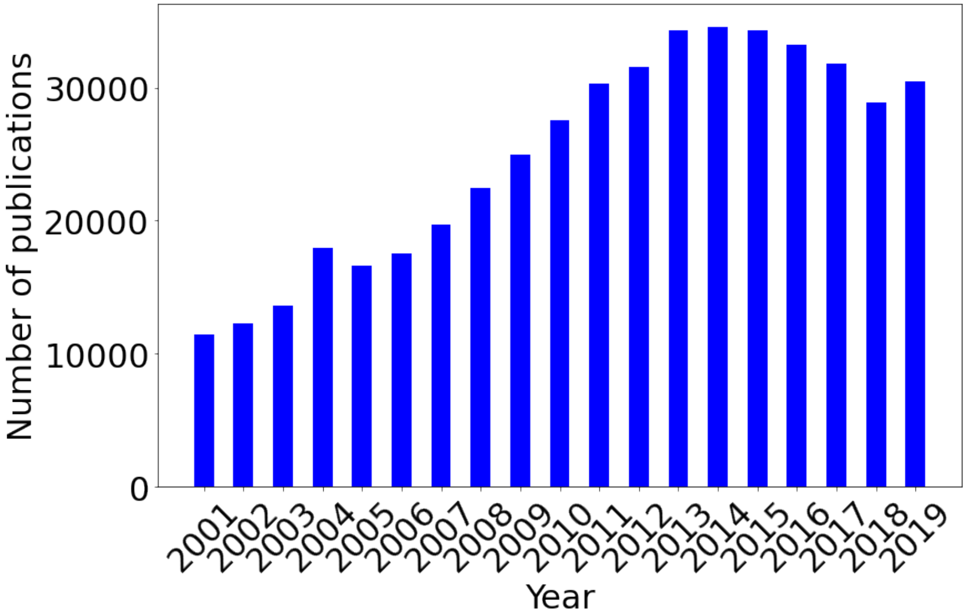
Joint work with Joshua Bergers, Duane Verner, LA Levy, John Hutchison, John Murphy
Prasanna Balaprakash, Yan Feng
Argonne National Laboratory

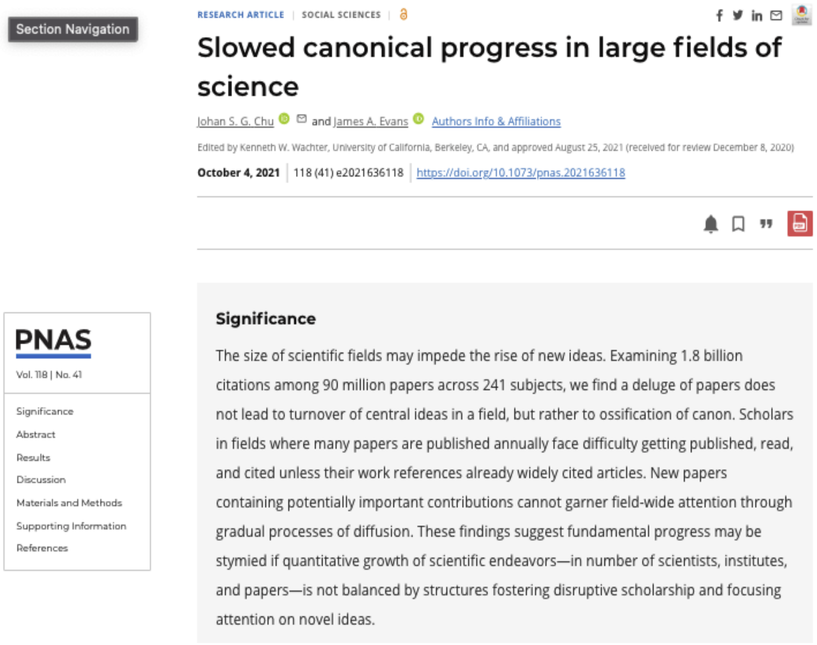
Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

IMPACT OF CLIMATE CHANGE ON CRITICAL INFRASTRUCTURE

CHALLENGE: ANALYZING LARGE CORPUS OF SCIENTIFIC LITERATURE

Growing number of publications





RESEARCH ARTICLE | SOCIAL SCIENCES |

Slowed canonical progress in large fields of science

Johan S. G. Chu and James A. Evans Authors Info & Affiliations

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved August 25, 2021 (received for review December 8, 2020)

October 4, 2021 | 118 (41) e2021636118 | https://doi.org/10.1073/pnas.2021636118

PNAS

Vol. 118 | No. 41

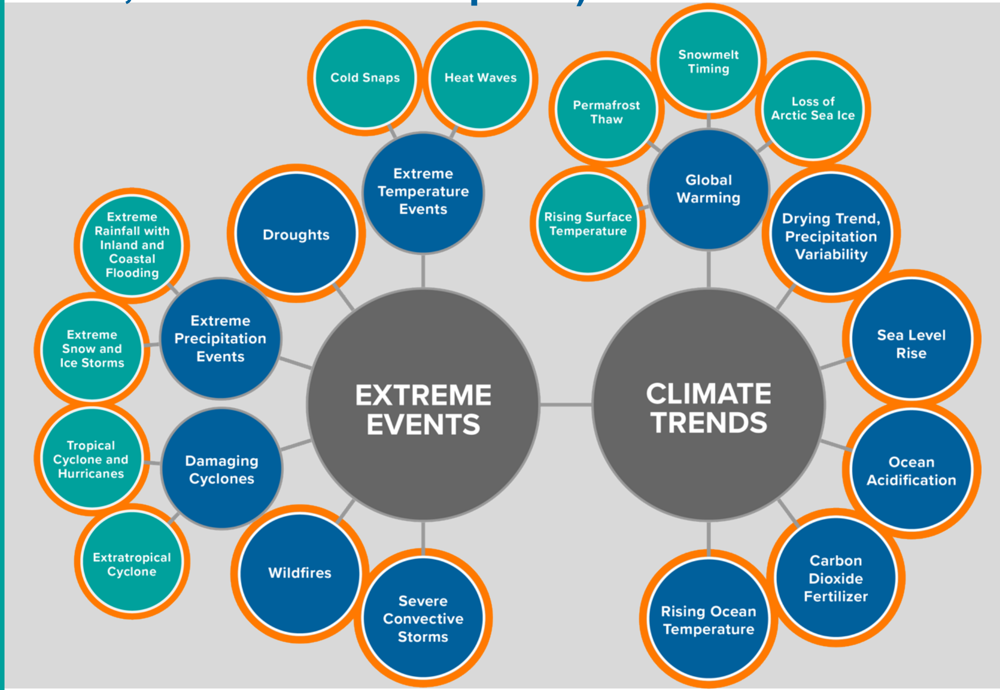
Significance
Abstract
Results
Discussion
Materials and Methods
Supporting Information
References

Significance

The size of scientific fields may impede the rise of new ideas. Examining 1.8 billion citations among 90 million papers across 241 subjects, we find a deluge of papers does not lead to turnover of central ideas in a field, but rather to ossification of canon. Scholars in fields where many papers are published annually face difficulty getting published, read, and cited unless their work references already widely cited articles. New papers containing potentially important contributions cannot garner field-wide attention through gradual processes of diffusion. These findings suggest fundamental progress may be stymied if quantitative growth of scientific endeavors—in number of scientists, institutes, and papers—is not balanced by structures fostering disruptive scholarship and focusing attention on novel ideas.

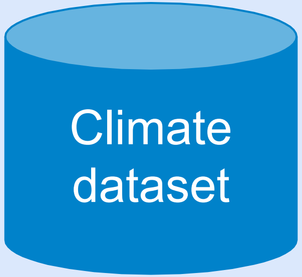
CLIMATE HAZARDS CATEGORIES AND DATASET

18 Climate Change Hazards (2021 IPCC, NASEM, ASCE, and USGCRP reports)


EXTREME EVENTS
Cold Snaps
Heat Waves
Extreme Temperature Events
Droughts
Extreme Rainfall with Inland and Coastal Flooding
Extreme Precipitation Events
Extreme Snow and Ice Storms
Tropical Cyclone and Hurricanes
Damaging Cyclones
Extratropical Cyclone
Wildfires
Severe Convective Storms
CLIMATE TRENDS
Snowmelt Timing
Permafrost Thaw
Loss of Arctic Sea Ice
Global Warming
Rising Surface Temperature
Drying Trend, Precipitation Variability
Sea Level Rise
Ocean Acidification
Carbon Dioxide Fertilizer
Rising Ocean Temperature

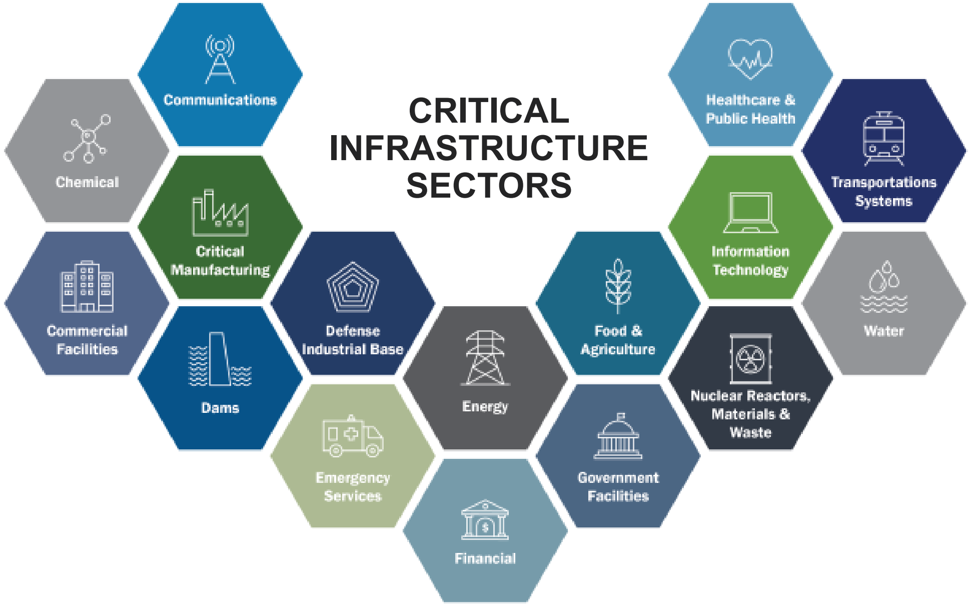
Unlabeled Dataset

- Subset Semantic Scholar Open Research Corpus (S2ORC) to develop climate change corpora
  - 600K climate-related articles



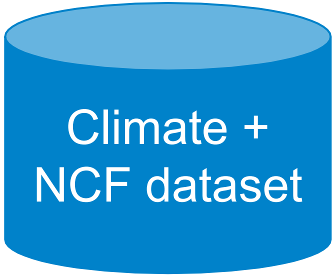
RESILIENCE AND SECURITY

Critical Infrastructure and 55 National Critical Functions



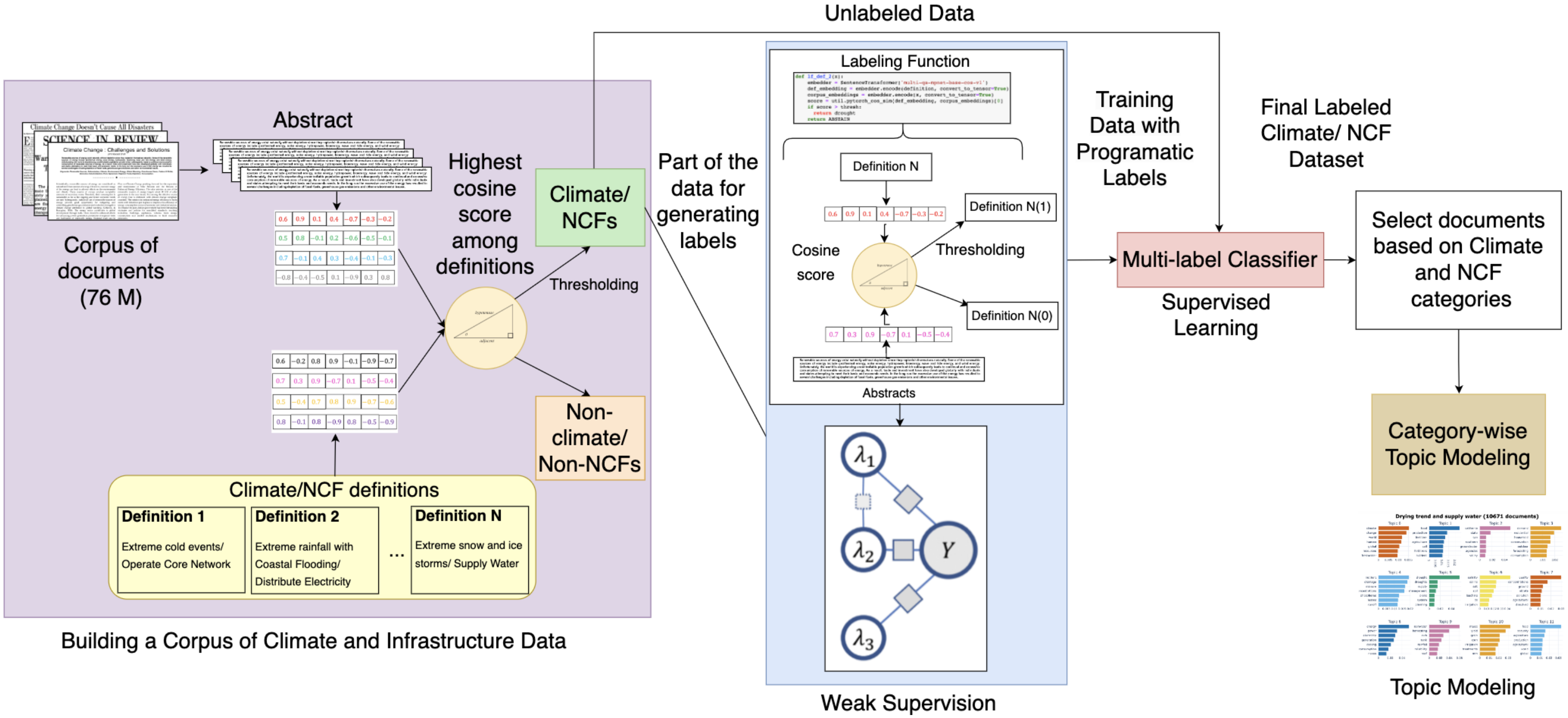
source: https://www.cisa.gov/identifying-critical-infrastructure-during-covid-19

- **Subset climate related corpus to develop climate and NCFs corpora**
  - **17K climate and NCFs related articles**



PIPELINE FOR LABELING LARGE CORPUS

Weak Supervision and Supervised learning



Unlabeled Data

Corpus of documents (76 M)

Abstract

0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

Highest cosine score among definitions

Thresholding

Climate/ NCFs

Non-climate/ Non-NCFs

0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Climate/NCF definitions

Definition 1
Extreme cold events/ Operate Core Network

Definition 2
Extreme rainfall with Coastal Flooding/ Distribute Electricity

…

Definition N
Extreme snow and ice storms/ Supply Water

Building a Corpus of Climate and Infrastructure Data

Part of the data for generating labels

Labeling Function

```
def lf_def_2(x):
    embedder = SentenceTransformer('multi-qa-mpnet-base-cos-v1')
    def_embedding = embedder.encode(definition, convert_to_tensor=True)
    corpus_embeddings = embedder.encode(x, convert_to_tensor=True)
    score = util.pytorch_cos_sim(def_embedding, corpus_embeddings)[0]
    if score > thresh:
      return drought
    return ABSTAIN
```

Definition N

0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2

Cosine score

Thresholding

Definition N(1)

Definition N(0)

0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4

Abstracts

Weak Supervision

Training Data with Programatic Labels

Multi-label Classifier

Supervised Learning

Final Labeled Climate/ NCF Dataset

Select documents based on Climate and NCF categories

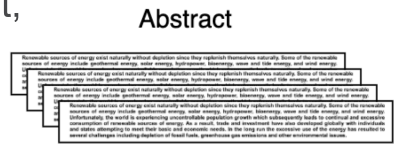
Category-wise Topic Modeling

Topic Modeling

WEAK SUPERVISION FOR GENERATING LABELS FOR TRAINING DATA

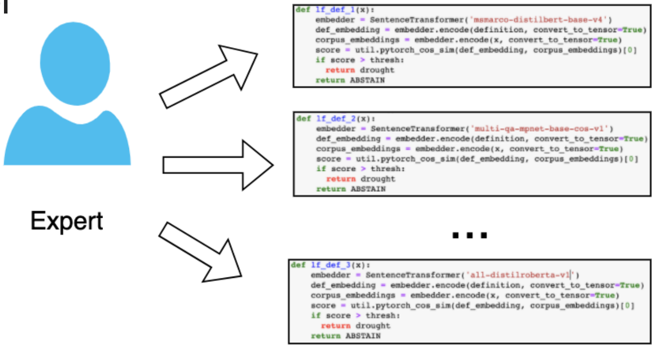
Weak Supervision using Snorkel to alleviate labeling bottleneck

- Labeling functions (LFs)
  - 7 LFs for 7 embedding techniques (distilbert, MiniLM, nli-stsb, mpnet, distilroberta, mpnet, msmarco)
  - Cosine similarity score
  - Optimization of overlapping and conflicts between LFs in weak supervised learning model



Abstract

Renewable sources of energy exist naturally without depletion since they replenish themselves naturally. Some of the renewable sources of energy include geothermal energy, solar energy, hydropower, bioenergy, wave and tide energy, and wind energy. Unfortunately, the world is experiencing uncontrollable population growth which subsequently leads to continual and excessive consumption of renewable sources of energy. As a result, trade and investment have also developed globally with individuals and states attempting to meet their basic and economic needs. In the long run the excessive use of the energy has resulted to several challenges including depletion of fossil fuels, greenhouse gas emissions and other environmental issues.



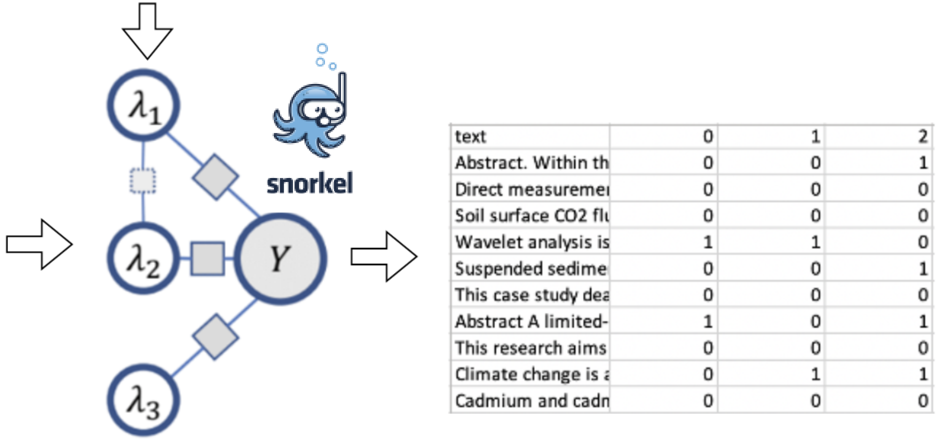
```
def lf_def_1(x):
    embedder = SentenceTransformer('msmarco-distilbert-base-v4')
    def_embedding = embedder.encode(definition, convert_to_tensor=True)
    corpus_embeddings = embedder.encode(x, convert_to_tensor=True)
    score = util.pytorch_cos_sim(def_embedding, corpus_embeddings)[0]
    if score > thresh:
      return drought
    return ABSTAIN
```

```
def lf_def_2(x):
    embedder = SentenceTransformer('multi-qa-mpnet-base-cos-v1')
    def_embedding = embedder.encode(definition, convert_to_tensor=True)
    corpus_embeddings = embedder.encode(x, convert_to_tensor=True)
    score = util.pytorch_cos_sim(def_embedding, corpus_embeddings)[0]
    if score > thresh:
      return drought
    return ABSTAIN
```

...

```
def lf_def_3(x):
    embedder = SentenceTransformer('all-distilroberta-v1')
    def_embedding = embedder.encode(definition, convert_to_tensor=True)
    corpus_embeddings = embedder.encode(x, convert_to_tensor=True)
    score = util.pytorch_cos_sim(def_embedding, corpus_embeddings)[0]
    if score > thresh:
      return drought
    return ABSTAIN
```

Labeling Functions



text	0	1	2
Abstract. Within th	0	0	1
Direct measuremei	0	0	0
Soil surface CO2 flu	0	0	0
Wavelet analysis is	1	1	0
Suspended sedime	0	0	1
This case study dea	0	0	0
Abstract A limited-	1	0	1
This research aims	0	0	0
Climate change is a	0	1	1
Cadmium and cadn	0	0	0

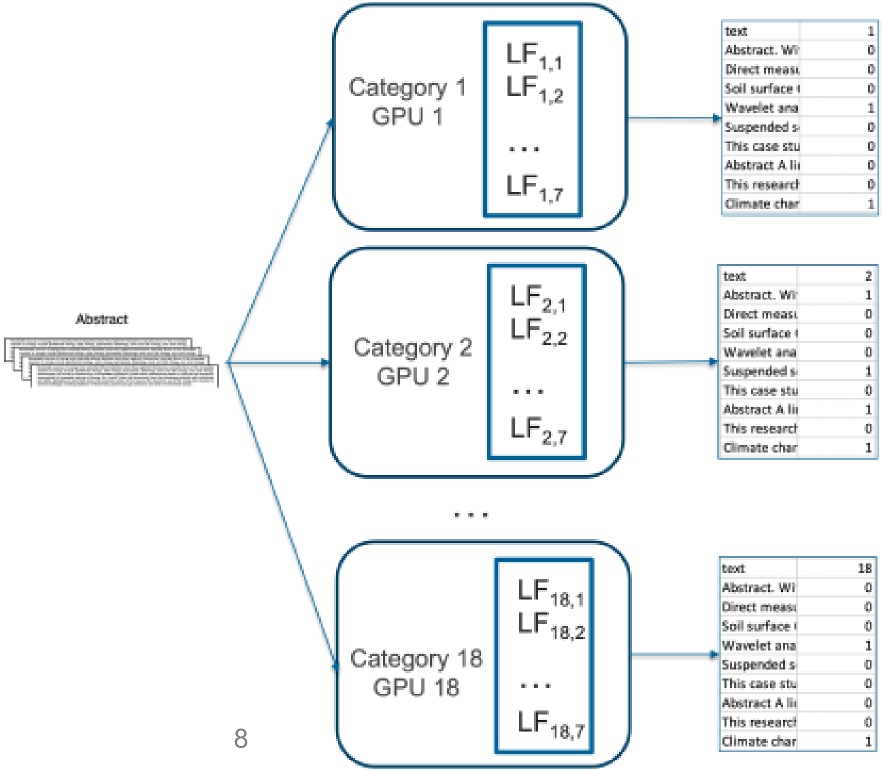
Weak supervised learning model

Labeled data

SCALING ON GPU CLUSTER

Programmatic rules for assigning labels to dataset

- Parallelization (GPU per class)
  - 18 GPUs labels 18 climate categories
  - Each GPU performs binary labeling (0/1)
  - Jobs run on multiple GPUs
- Parallelization (GPU per class)
  - 55 GPUs labels 55 NCF categories
  - Each GPU performs binary labeling (0/1)
  - Jobs run on multiple GPUs
- Labeled documents:
  - **~5K**



text	1
Abstract. Wi	0
Direct measu	0
Soil surface	0
Wavelet ana	1
Suspended s	0
This case stu	0
Abstract A li	0
This research	0
Climate char	1

text	2
Abstract. Wi	1
Direct measu	0
Soil surface	0
Wavelet ana	0
Suspended s	1
This case stu	0
Abstract A li	1
This research	0
Climate char	1

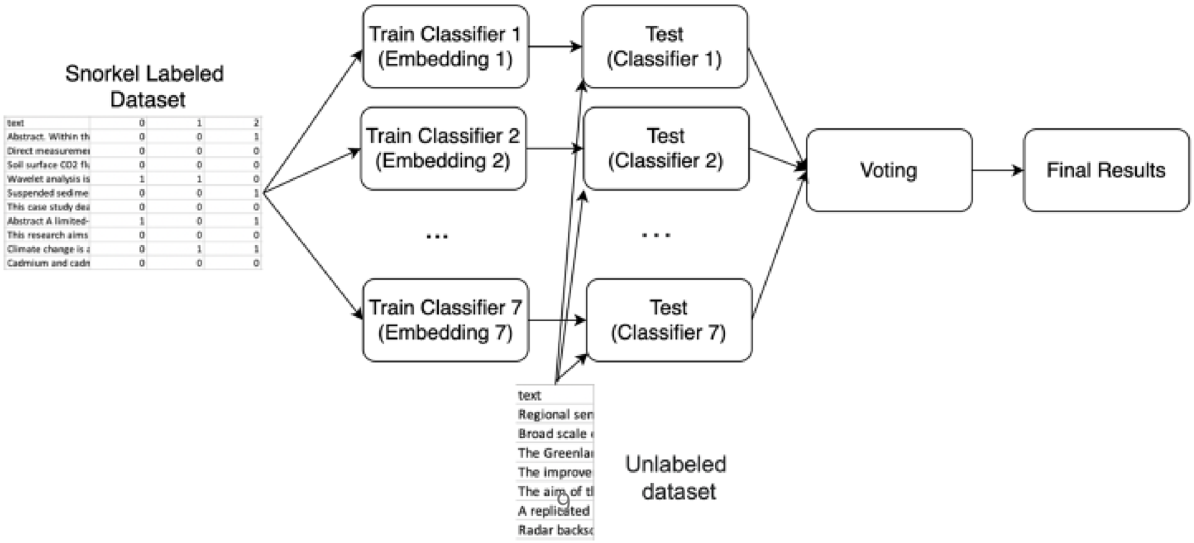
text	18
Abstract. Wi	0
Direct measu	0
Soil surface	0
Wavelet ana	1
Suspended s	0
This case stu	0
Abstract A li	0
This research	0
Climate char	1

SUPERVISED LEARNING

Multilabel classification using classifier chain

- Training data: ~5k (climate/NCFs)
- Test data: ~12k
- Model: Classifier chain with MLP
- Train 7 models using 7 embedding techniques

- Train and test 7 models simultaneously
- Run time: 5 hours (approx.)
- Cross validation accuracy on train dataset: 91%



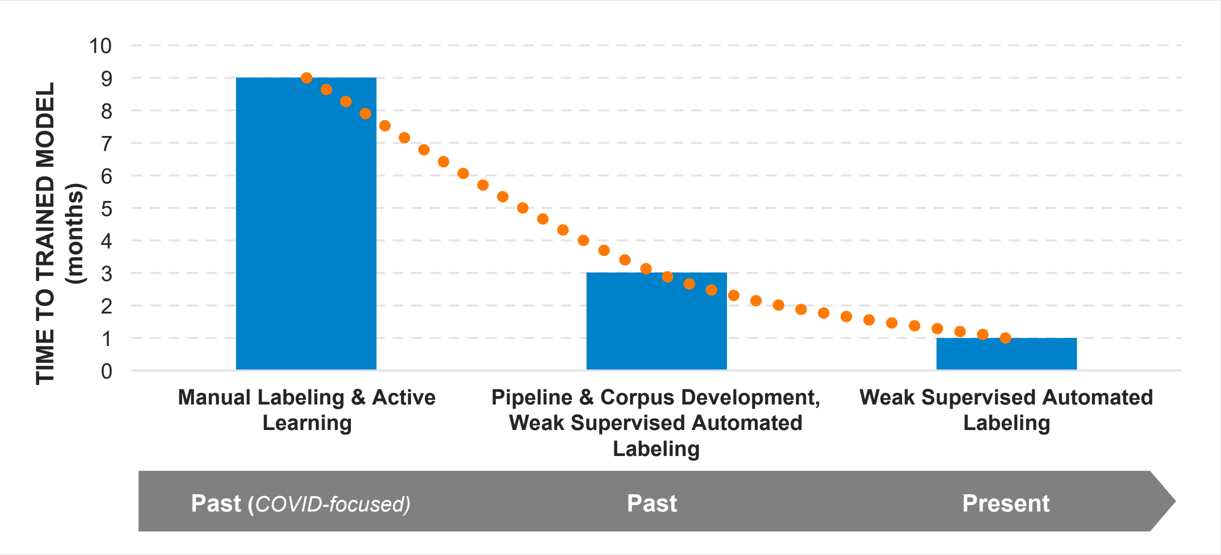
text	0	1	2
Abstract. Within th	0	0	1
Direct measureme	0	0	0
Soil surface CO2 fl	0	0	0
Wavelet analysis is	1	1	0
Suspended sedime	0	0	1
This case study dea	0	0	0
Abstract A limited-	1	0	1
This research aims	0	0	0
Climate change is a	0	1	1
Cadmium and cadn	0	0	0

text
Regional sen
Broad scale c
The Greenlar
The improve
The aim of th
A replicated
Radar backsc

GAINING EFFICIENCIES WITH WEAK SUPERVISED LEARNING

Expediting development of trained model

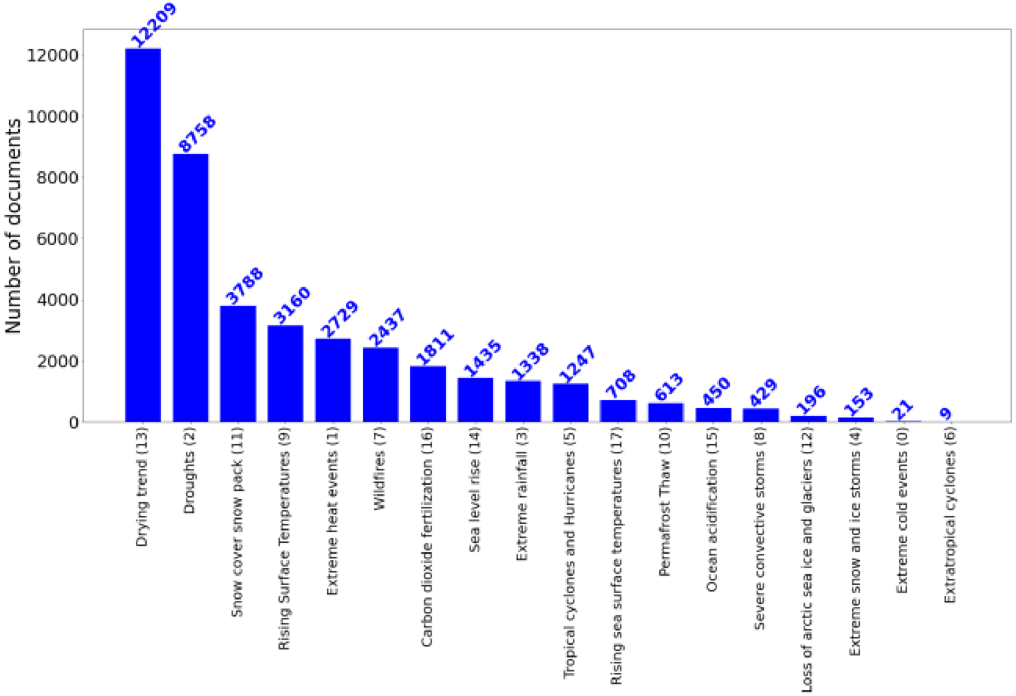
Automating the document labeling process addresses known “pain point” and enables fast, repeatable scaling



NUMBER OF DOCUMENTS PER CLIMATE CATEGORY

The trend shows warming events are more discussed than the cooling events

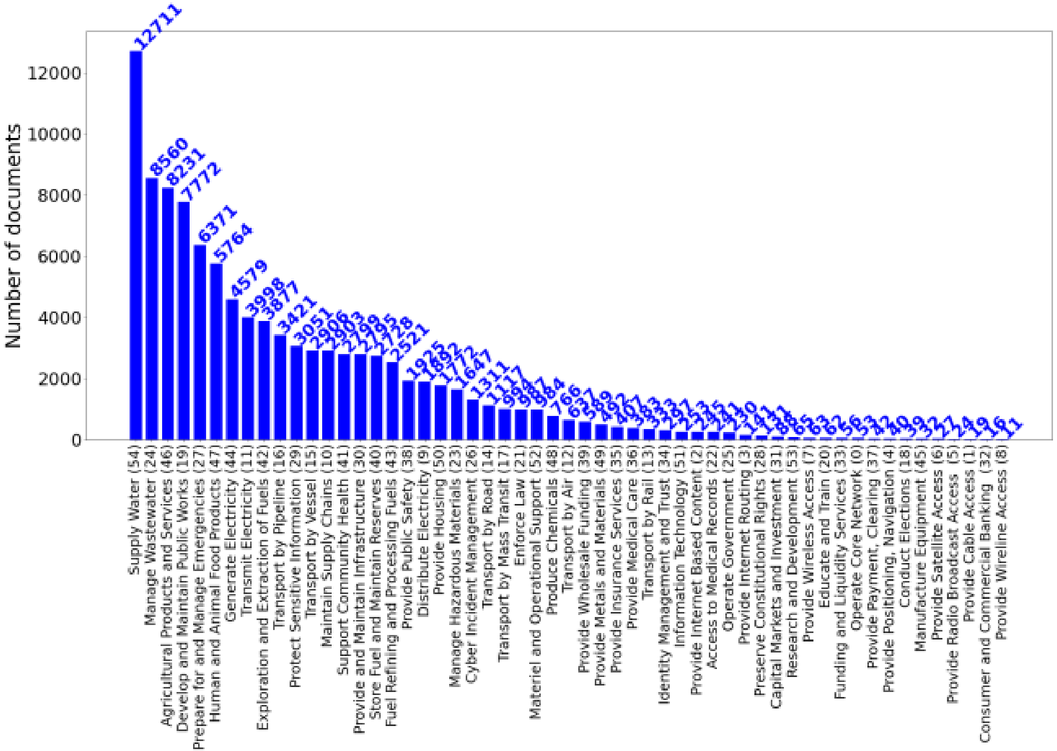
- Labeling the large corpus assisted us in identifying a recent trend in climate research
- Currently heat events like Drying trend, Rising surface temperature, Drought are the major concern



NUMBER OF DOCUMENTS PER NCF CATEGORY

The trend shows water related events are more discussed

- Labeling the large corpus assisted us in identifying a recent trend in NCF
- Currently supply and mange water are the major concern



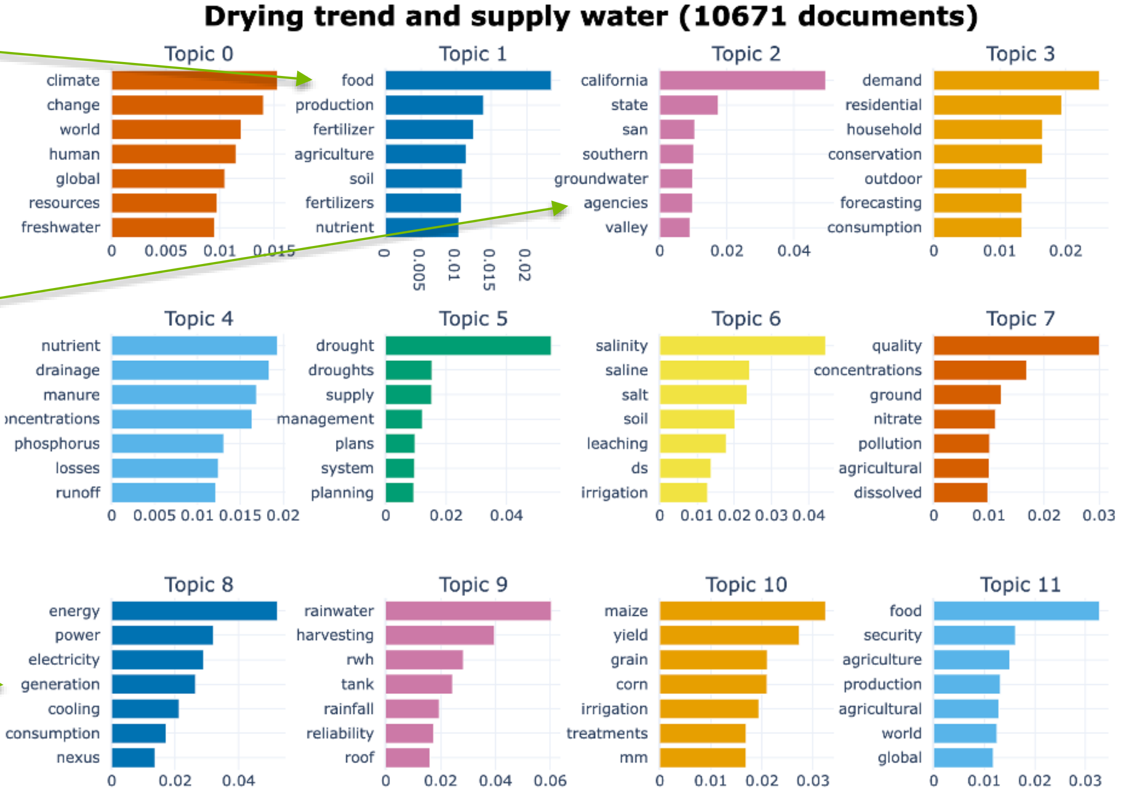
TOPIC MODELING

BERT based topic modeling on Climate hazard and NCF pairs

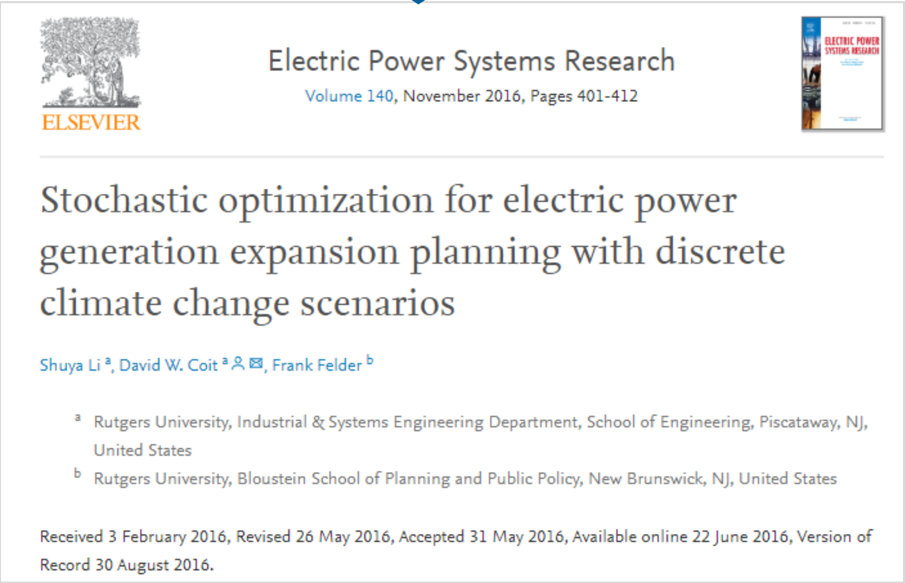
Drying trends and supplying water in the context of agriculture

Challenges associated with managing water supplies in southern California in the face of drying trends and historic droughts

Impacts of water availability on electric power generation technologies



If a researcher/analyst is looking at this article included in our corpus….



Electric Power Systems Research

Volume 140, November 2016, Pages 401-412

Stochastic optimization for electric power generation expansion planning with discrete climate change scenarios

Shuya Li [a], David W. Coit [a], Frank Felder [b]

[a] Rutgers University, Industrial & Systems Engineering Department, School of Engineering, Piscataway, NJ, United States

[b] Rutgers University, Bloustein School of Planning and Public Policy, New Brunswick, NJ, United States

Received 3 February 2016, Revised 26 May 2016, Accepted 31 May 2016, Available online 22 June 2016, Version of Record 30 August 2016.

…then some example next steps for exploration might include:

- Identify which climate hazards and NCFs this article is most frequently associated with
- Explore which topics for hazards/NCFs this article is most frequently associated with

ACKNOWLEDGMENT

Argonne
NATIONAL LABORATORY

LDRD
Laboratory Directed Research and Development

Climate LDRD Pillar 3

Argonne Leadership Computing Facility

Hugging Face

snorkel

Argonne
NATIONAL LABORATORY

THANK YOU

tmallick@anl.gov | https://tanwimallick.github.io/

Joint work with Joshua Bergers, Duane Verner, LA Levy, John Hutchison, John Murphy, Prasanna Balaprakash, Yan Feng