Using Machine Learning to Classify Earth Science Articles

Andrew Cramer, NASA Summer 2021 Intern, Texas A&M University
Vanessa Chatman, NASA Summer 2021 Intern, Northeastern University

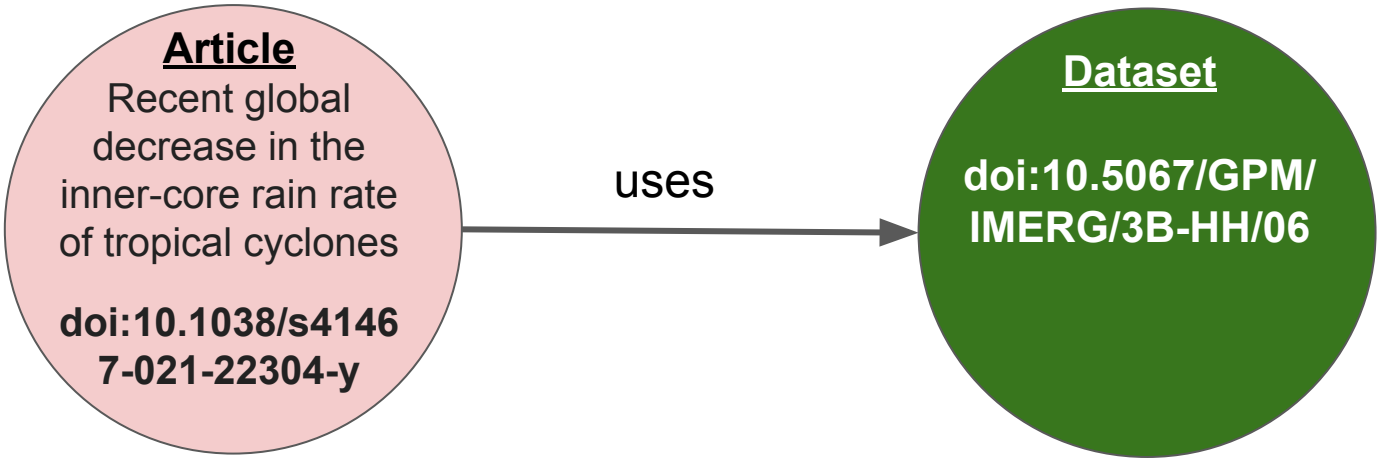
Acknowledgements

- NASA IMPACT Project
- Irina Gerasimov
- Dr. Elena Steponaitis
- Astrophysics Data System
- ESIP Discovery Cluster
- ESIP Foraging Participants
- Dr. Chris Lynnes
- Vincent Inverso

IMPACT = Interagency Implementation and Advanced Concepts Team

ESIP = Federation of Earth Science Information Partners

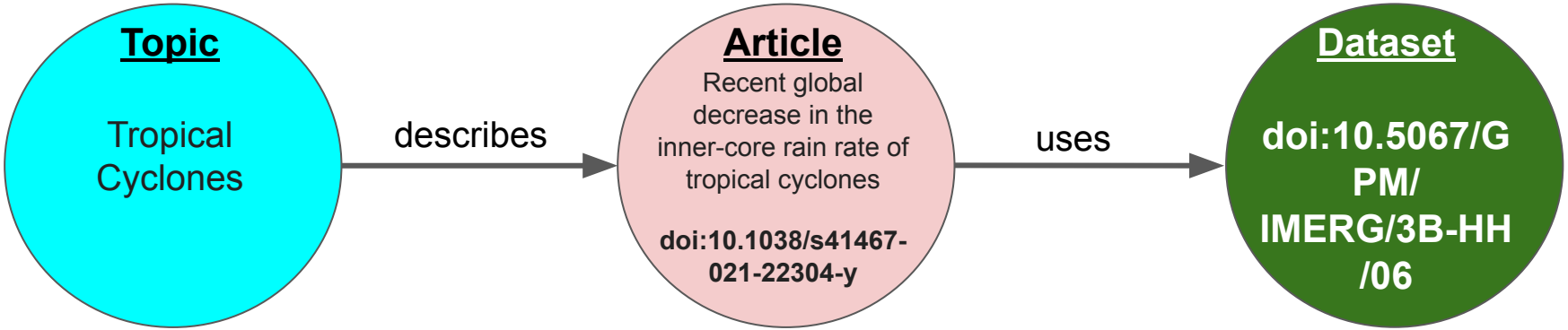
The Usage-Based Discovery (UBD) tool provides a linkage between a a research article and its Earth science dataset(s).



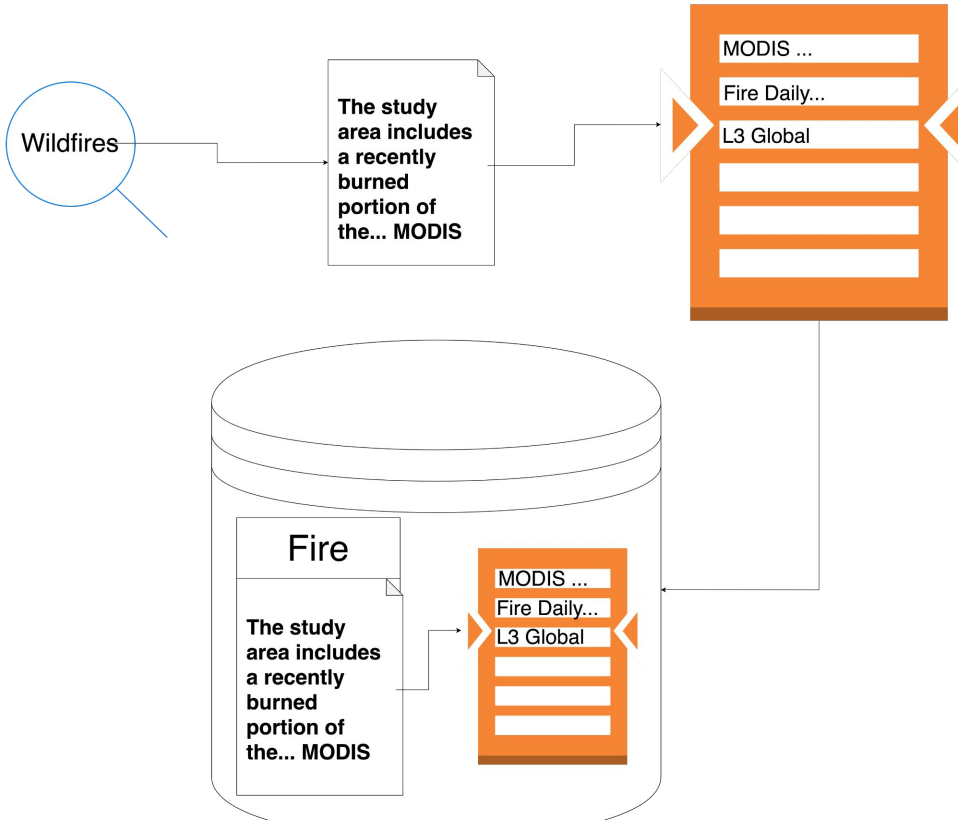
Benefits

- Links Earth science articles and their datasets even when citation is lacking
- Provides information on the usage cases and breadth of NASA data products
- Recommends helpful datasets for an end-user researching a data topic

Populating the UBD required a confirmed article topic, the article itself, and its used dataset(s).

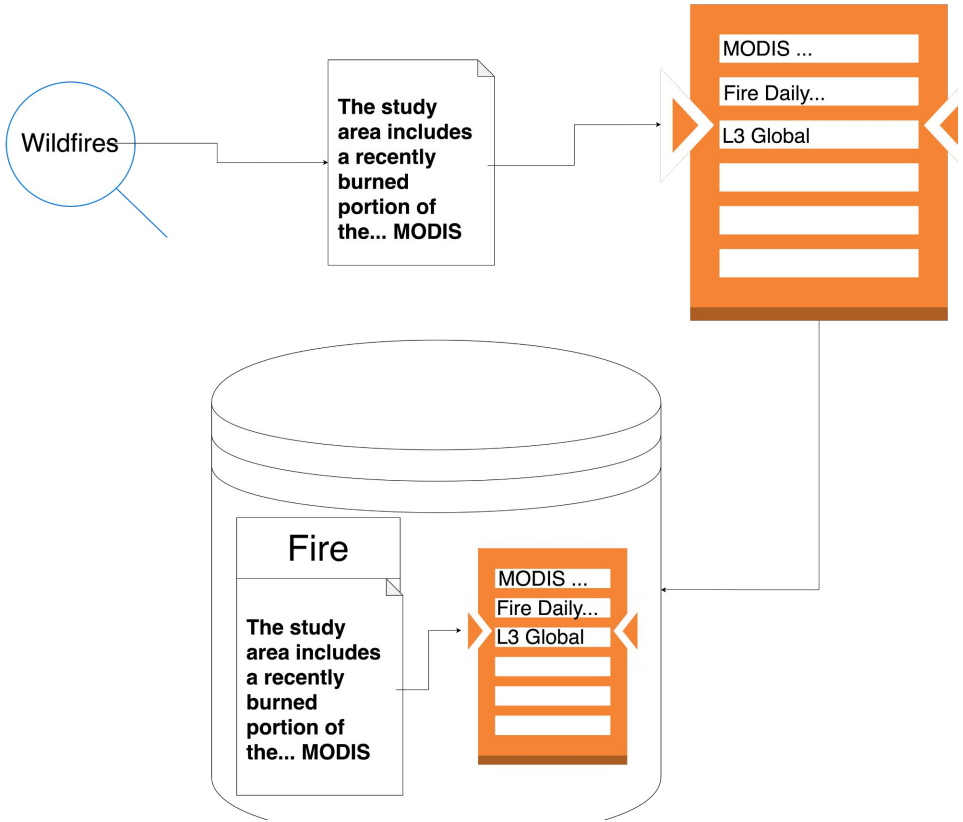


The initial UBD population pipeline follows manual steps:



1. Search for the article by topic.
2. Scan through the article for dataset reference.
3. Once found, confirm the article topic for later navigability.
4. Lastly, add this article-dataset linkage to UBD database.

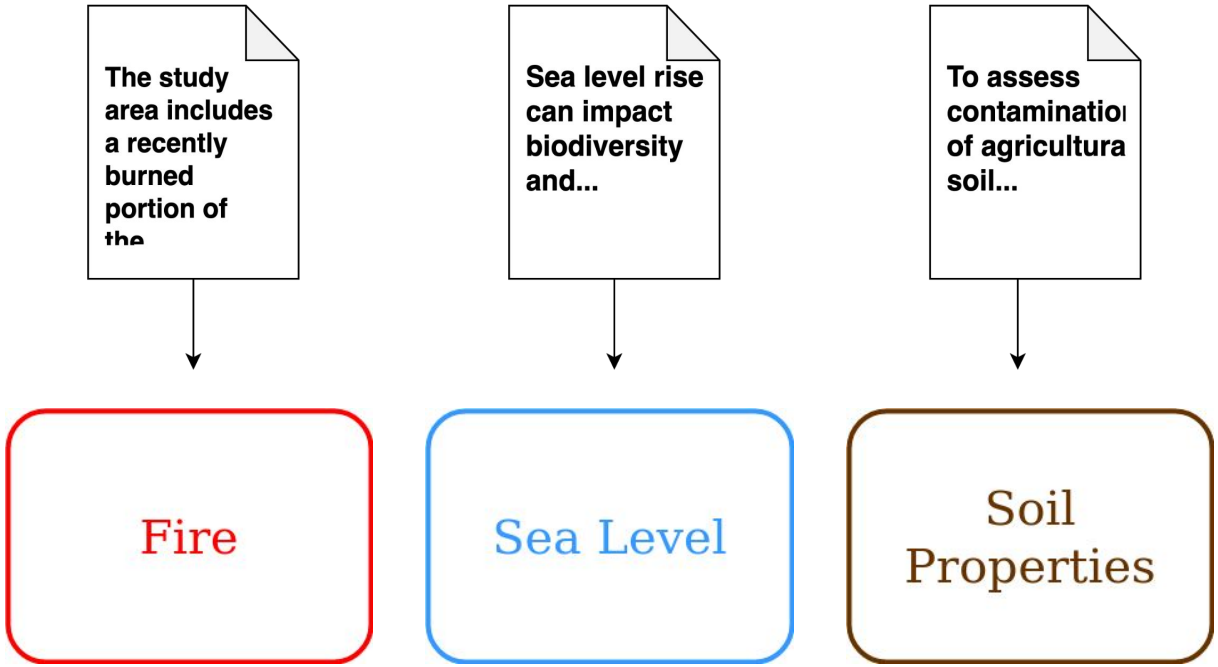
The initial UBD population pipeline follows manual steps:



3. Once found, confirm the article topic for later navigability.

This task proves to be labor intensive and low-yielding.

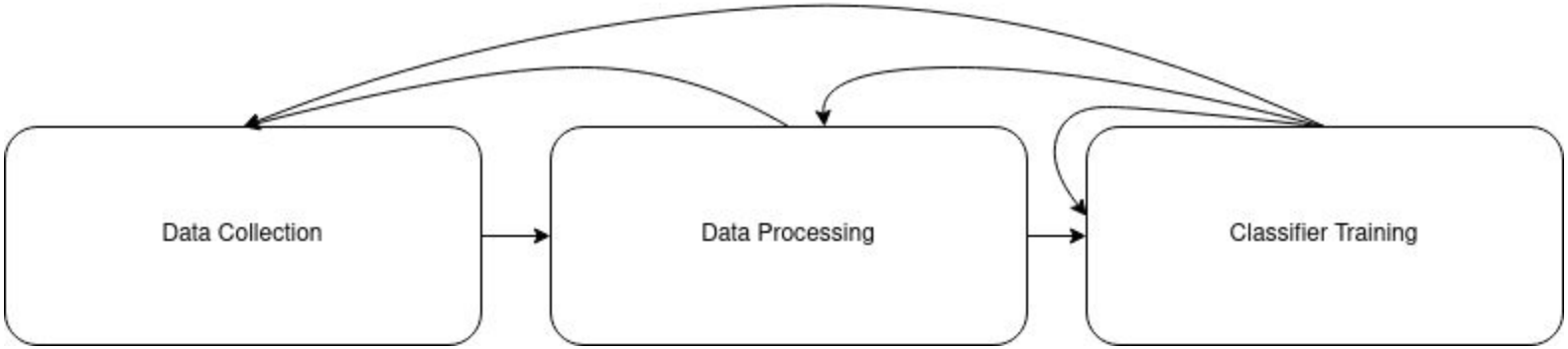
Idea: Topic classification can be automated with machine learning.



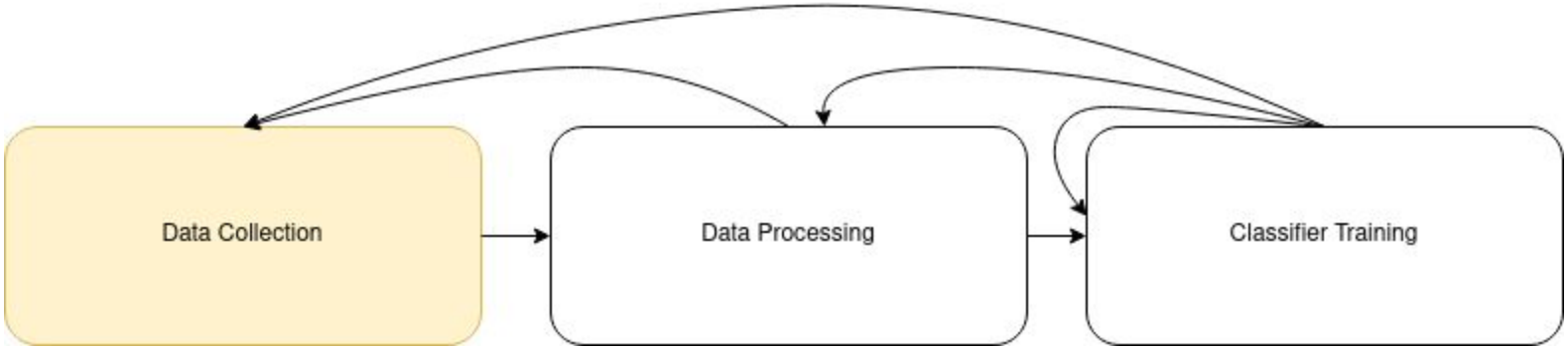
Five basic steps to utilizing machine learning in a new space:

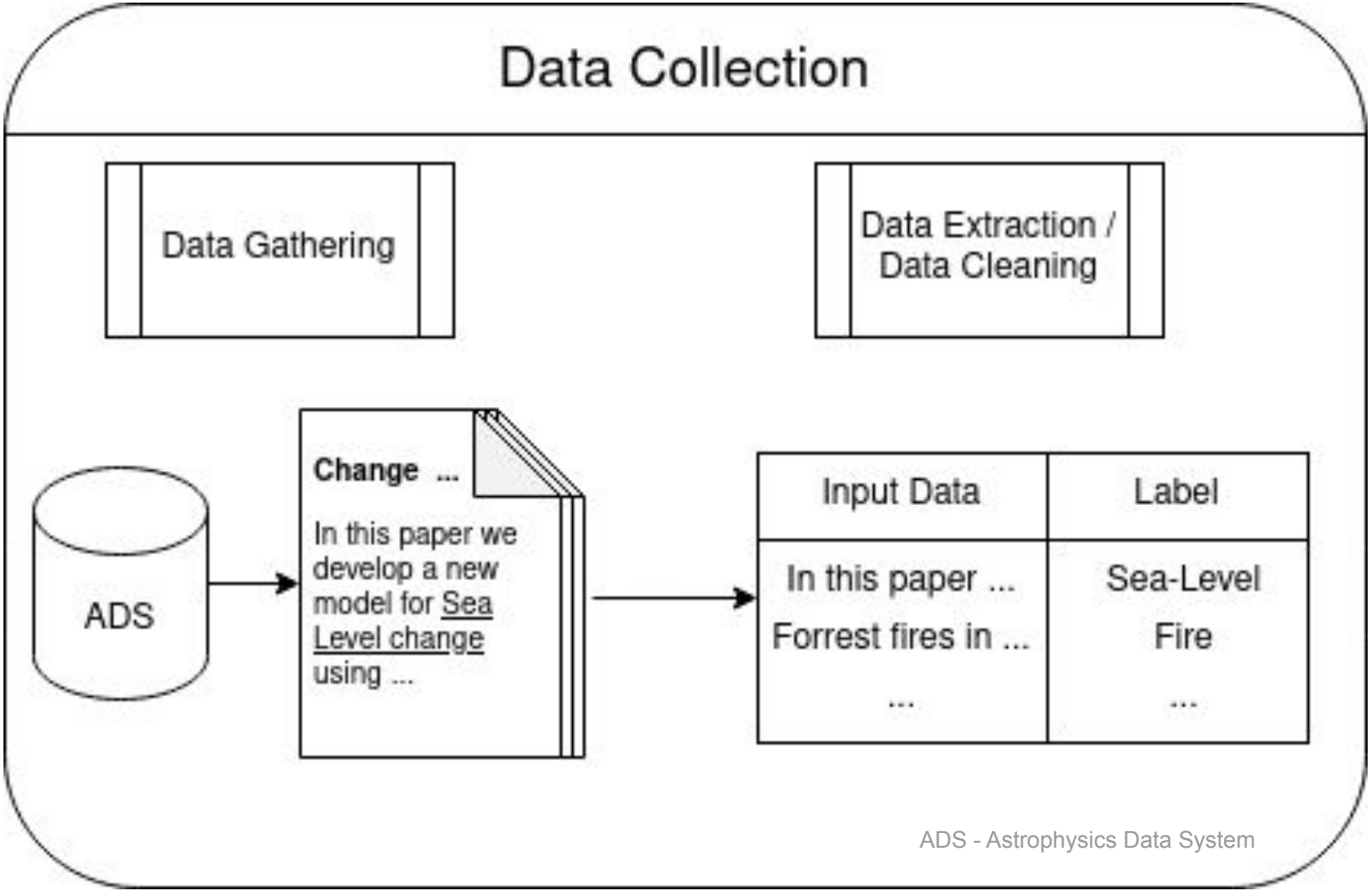
1. Define task
2. Collect data
3. Train Classifier
4. Evaluate
5. Repeat

Our machine learning pipeline development process:



collects research articles, **processes** (transforms) them into a format readable by a machine, and **classifies** them under a topic using a classification algorithm.

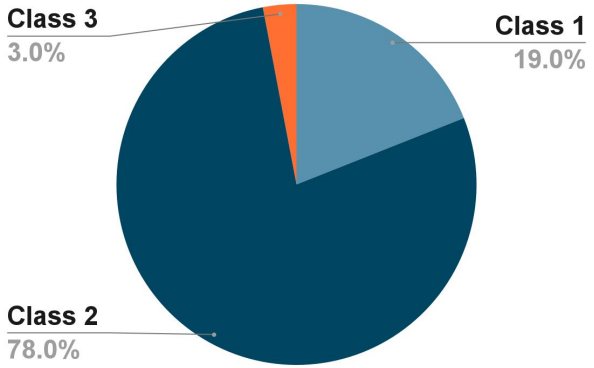

Data Collection
Data Processing
Classifier Training



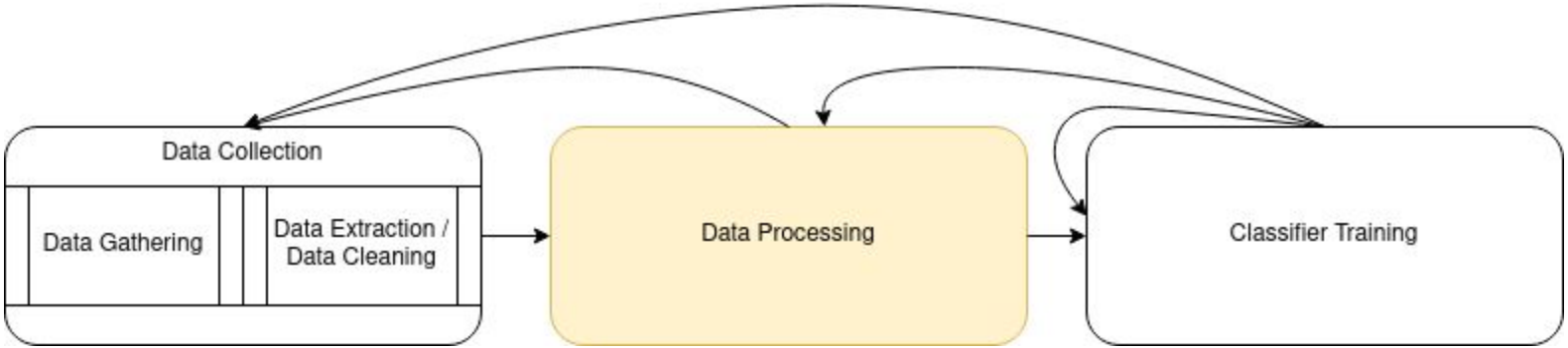
ADS - Astrophysics Data System

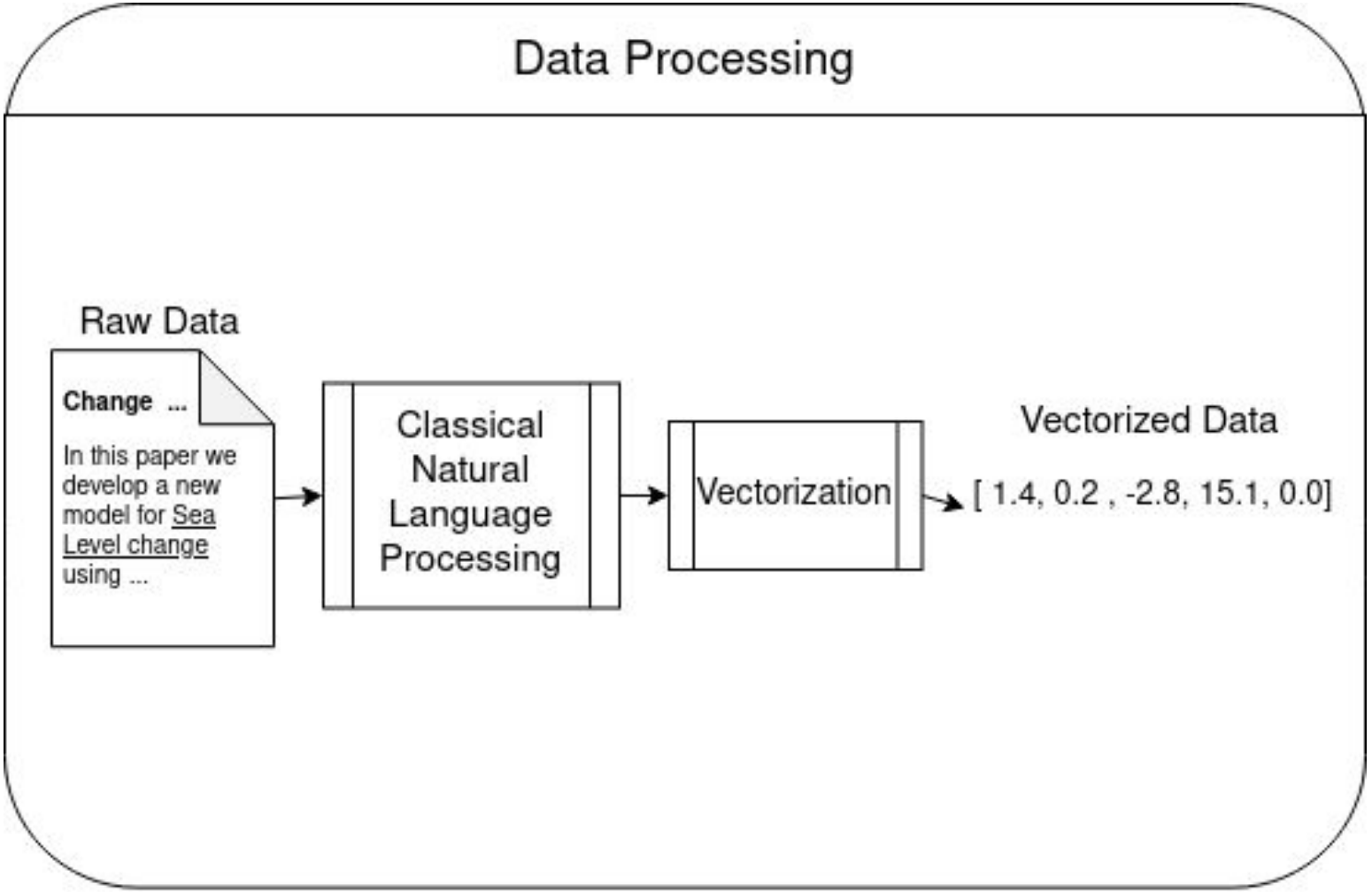
What we learned to be important in Dataset Collection:

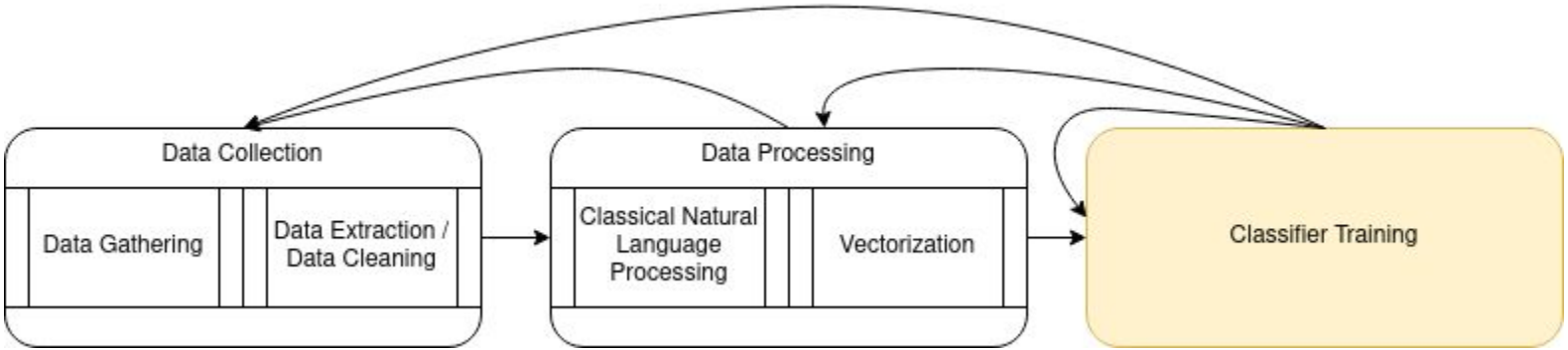
- Distribution and bias

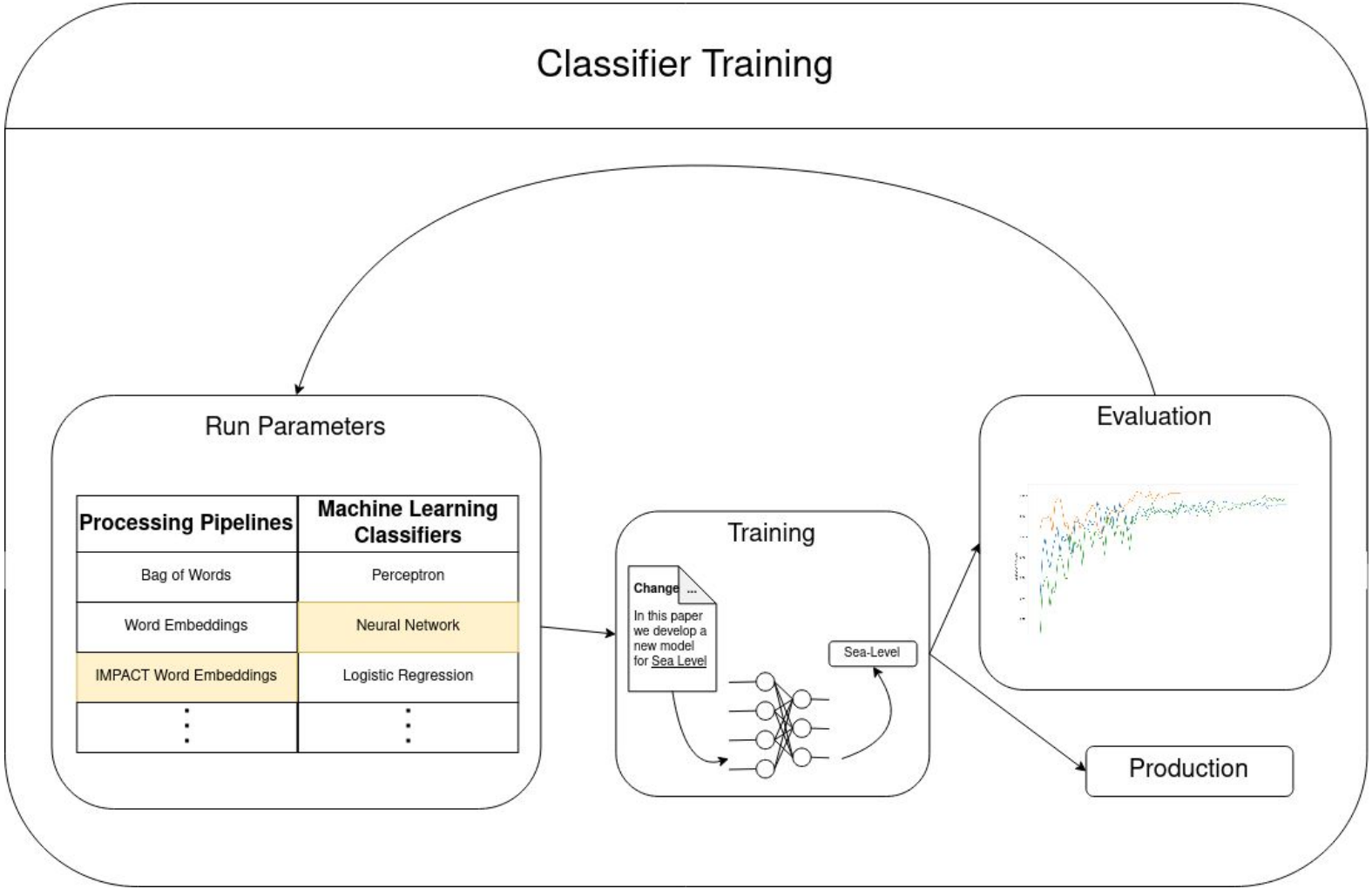


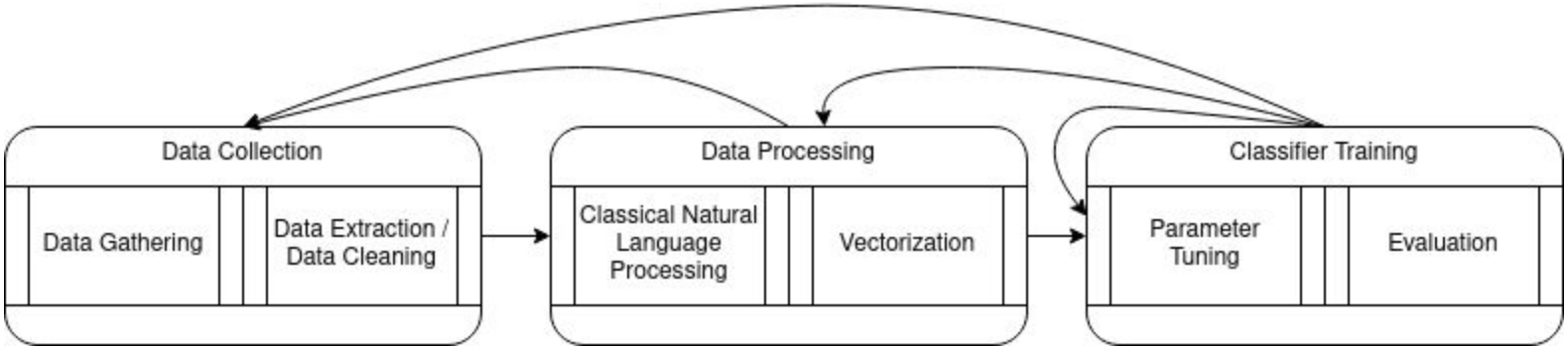
- Accuracy vs breadth and size


Data Collection
Data Gathering
Data Extraction / Data Cleaning
Data Processing
Classifier Training


Data Processing
Raw Data
Change ...
In this paper we develop a new model for Sea Level change using ...
Classical Natural Language Processing
Vectorization
Vectorized Data
[1.4, 0.2 , -2.8, 15.1, 0.0]


Data Collection
Data Gathering
Data Extraction / Data Cleaning
Data Processing
Classical Natural Language Processing
Vectorization
Classifier Training


Classifier Training
Run Parameters
Processing Pipelines
Machine Learning Classifiers
Bag of Words
Perceptron
Word Embeddings
Neural Network
IMPACT Word Embeddings
Logistic Regression
Training
Change ...
In this paper we develop a new model for Sea Level
Sea-Level
Evaluation
Production


Data Collection
Data Gathering
Data Extraction / Data Cleaning
Data Processing
Classical Natural Language Processing
Vectorization
Classifier Training
Parameter Tuning
Evaluation

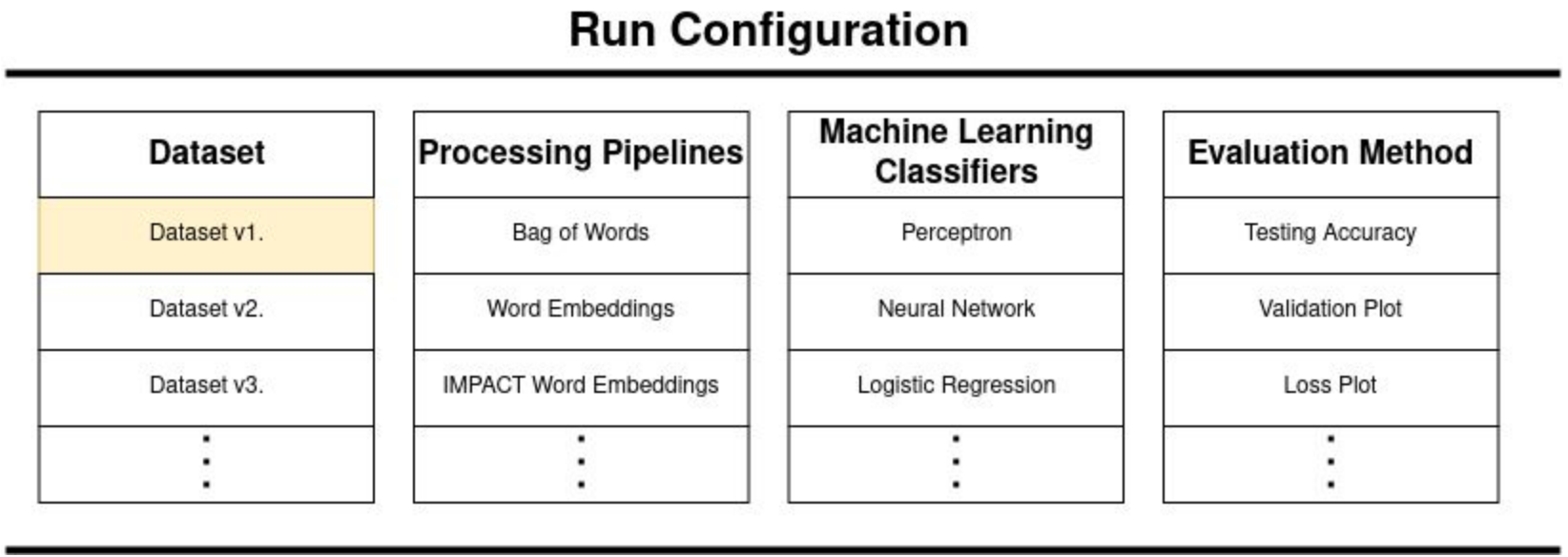
Design Tip: Prototyping various classification approaches

Run Configuration

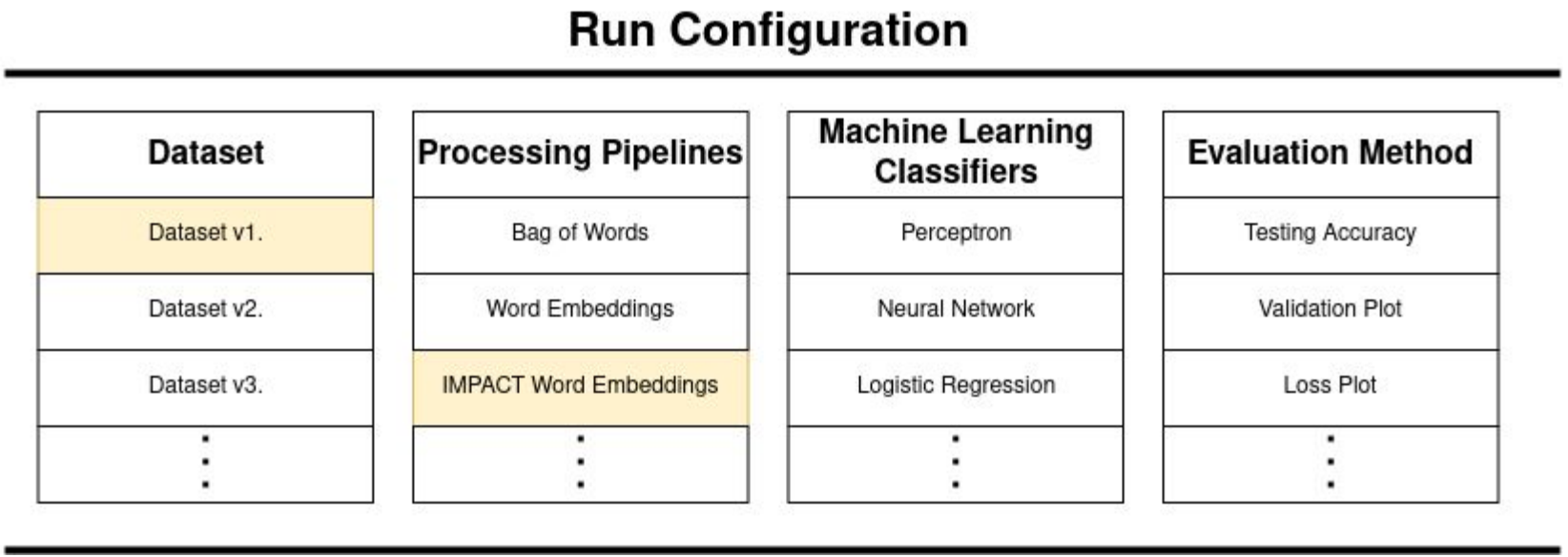
Dataset	Processing Pipelines	Machine Learning Classifiers	Evaluation Method
Dataset v1.	Bag of Words	Perceptron	Testing Accuracy
Dataset v2.	Word Embeddings	Neural Network	Validation Plot
Dataset v3.	IMPACT Word Embeddings	Logistic Regression	Loss Plot
⋮	⋮	⋮	⋮

The best classification results are initially unknown. Prototyping allows experimentation with different datasets, processing methods, classifiers, and result visualizations.

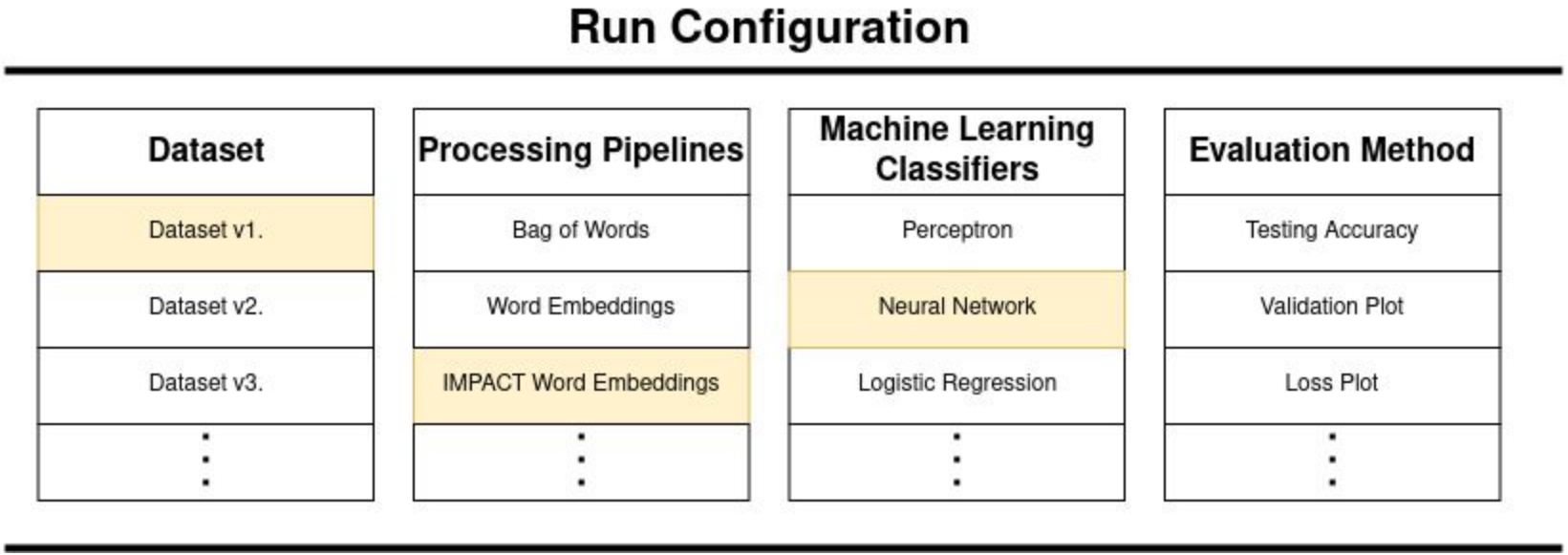
Design Tip: Prototyping



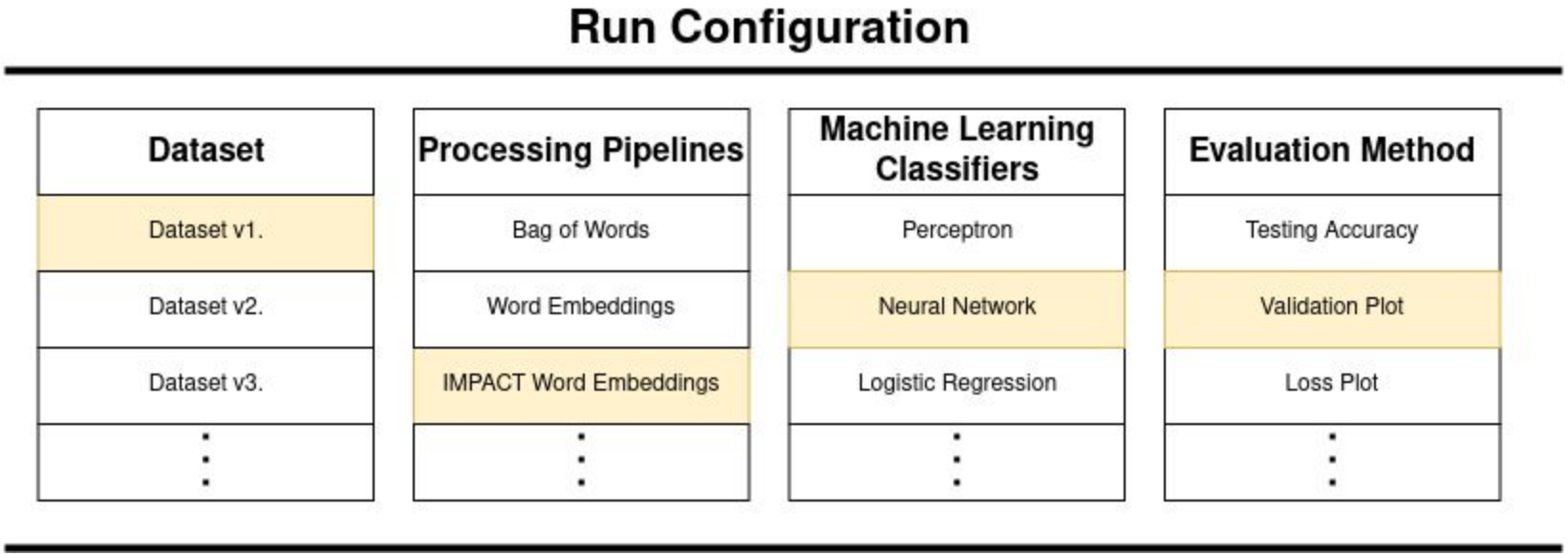
Design Tip: Prototyping



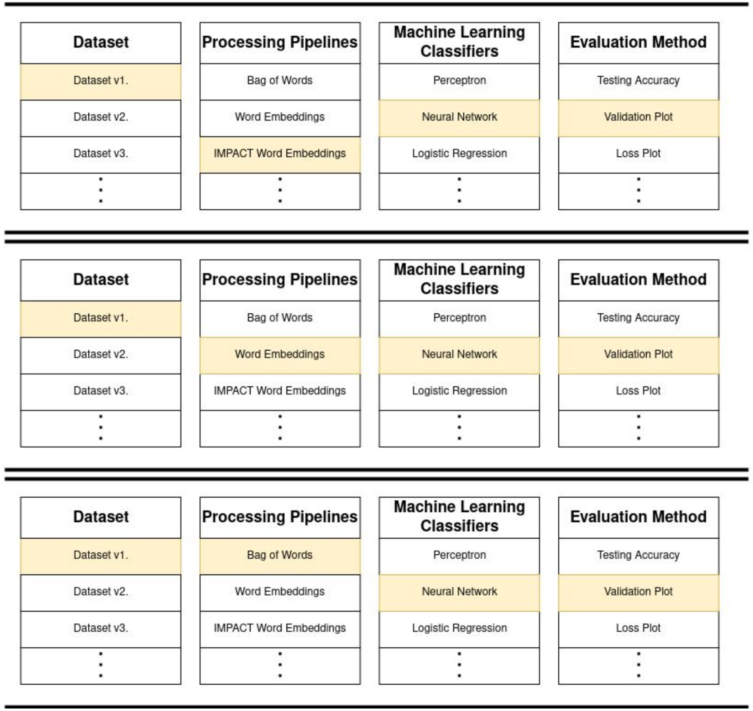
Design Tip: Prototyping

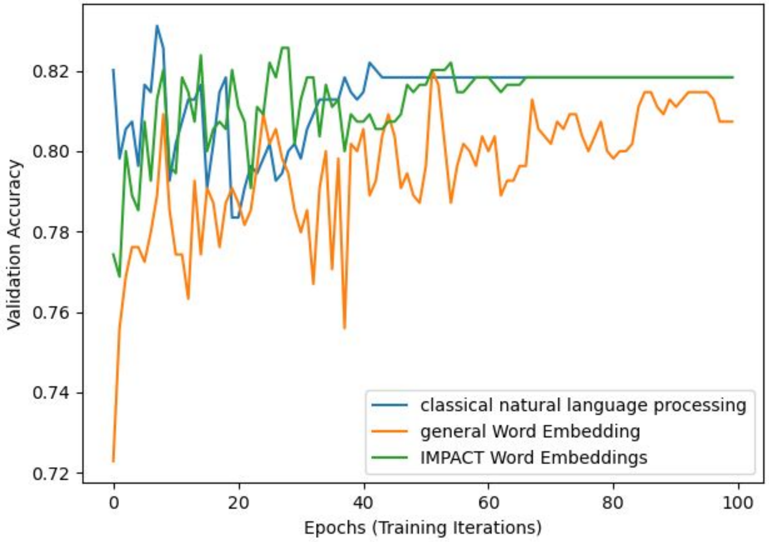


Design Tip: Prototyping

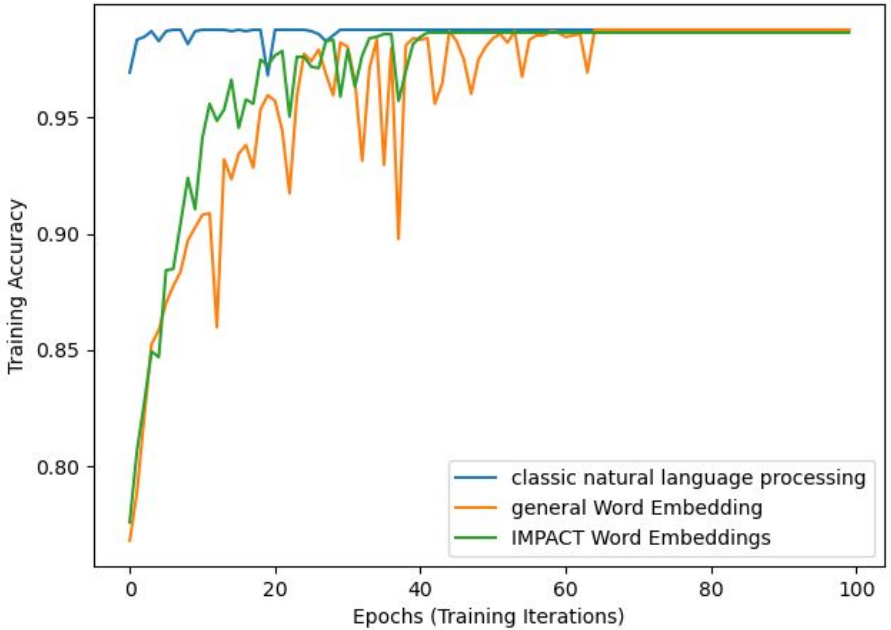


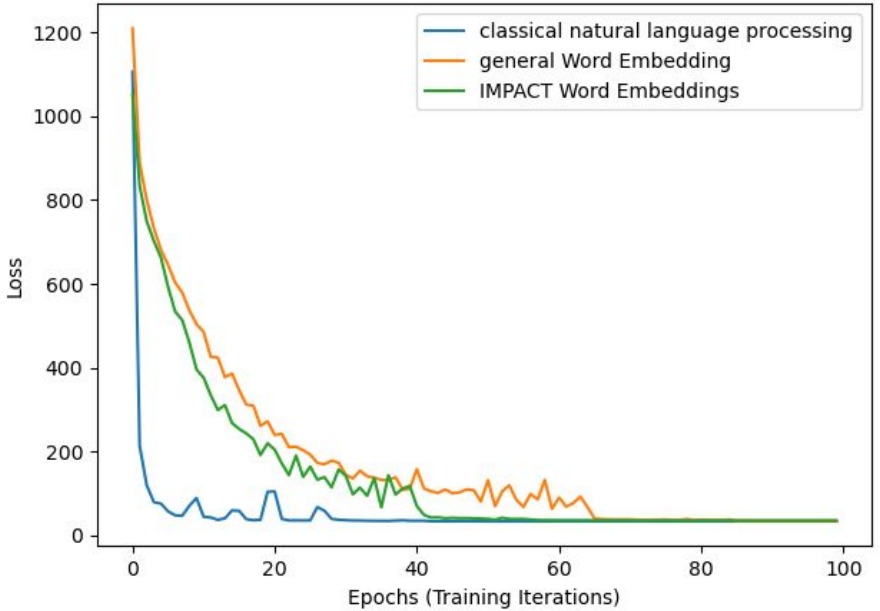
Design Tip: Prototyping



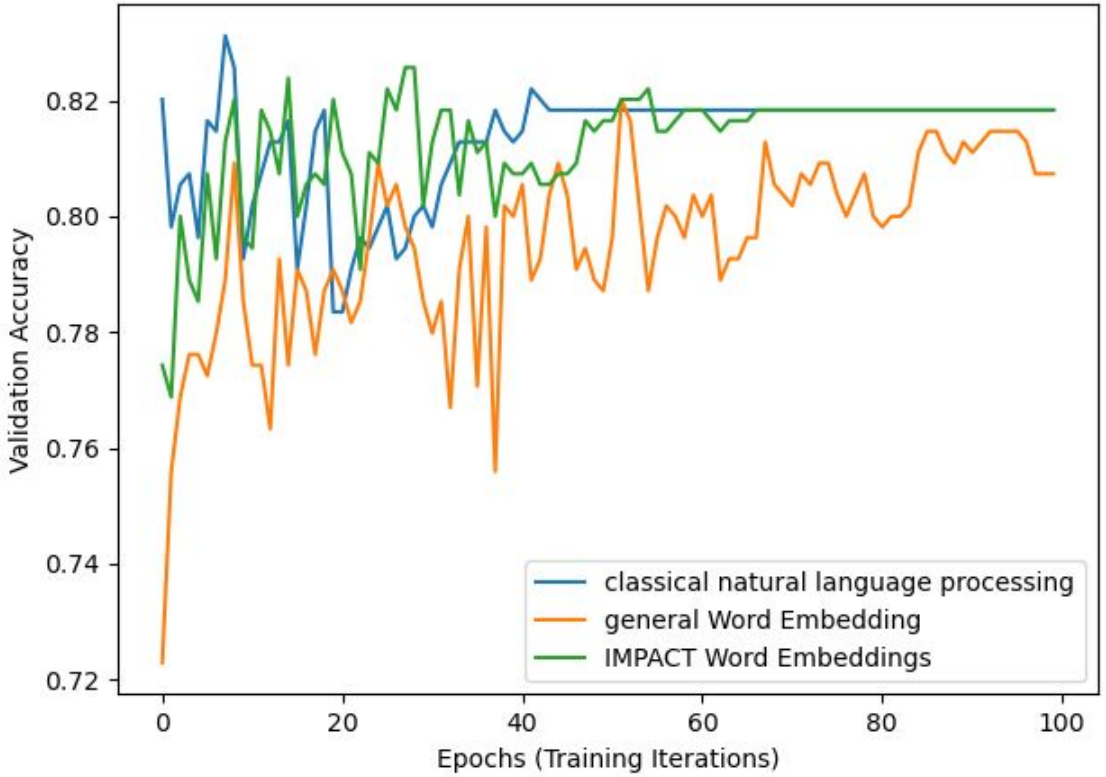


Evaluation Plotting

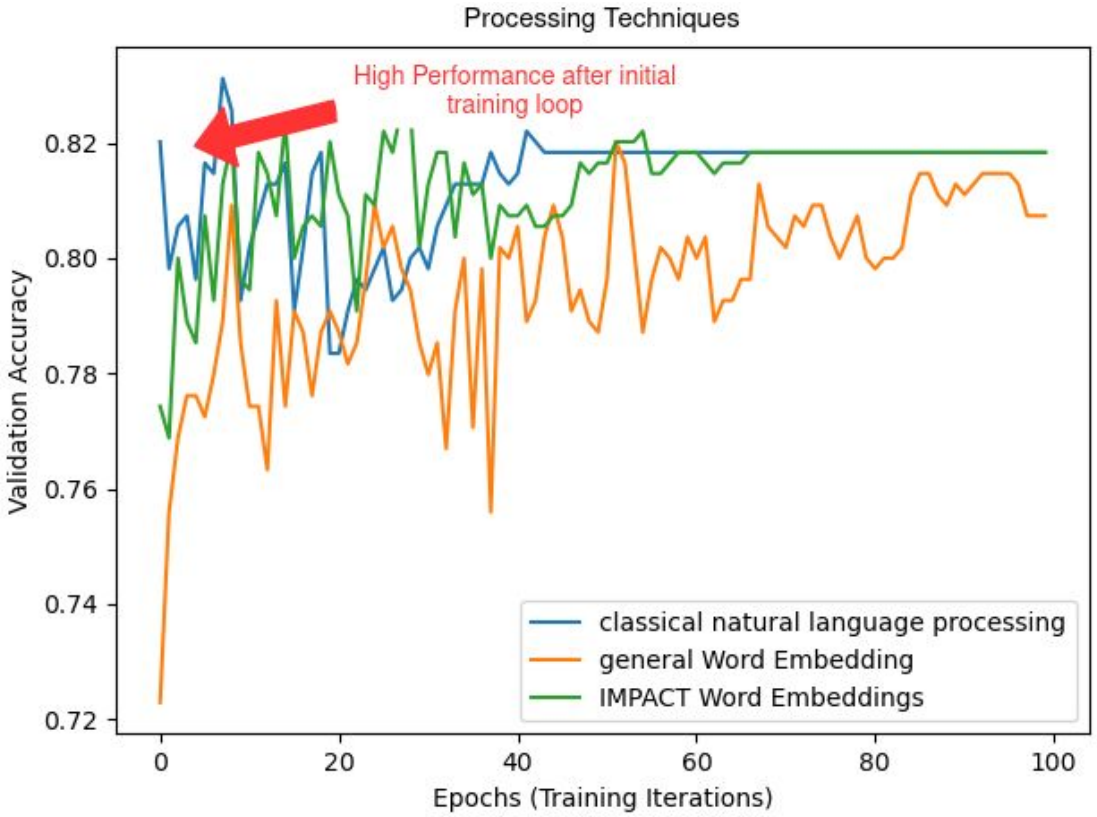




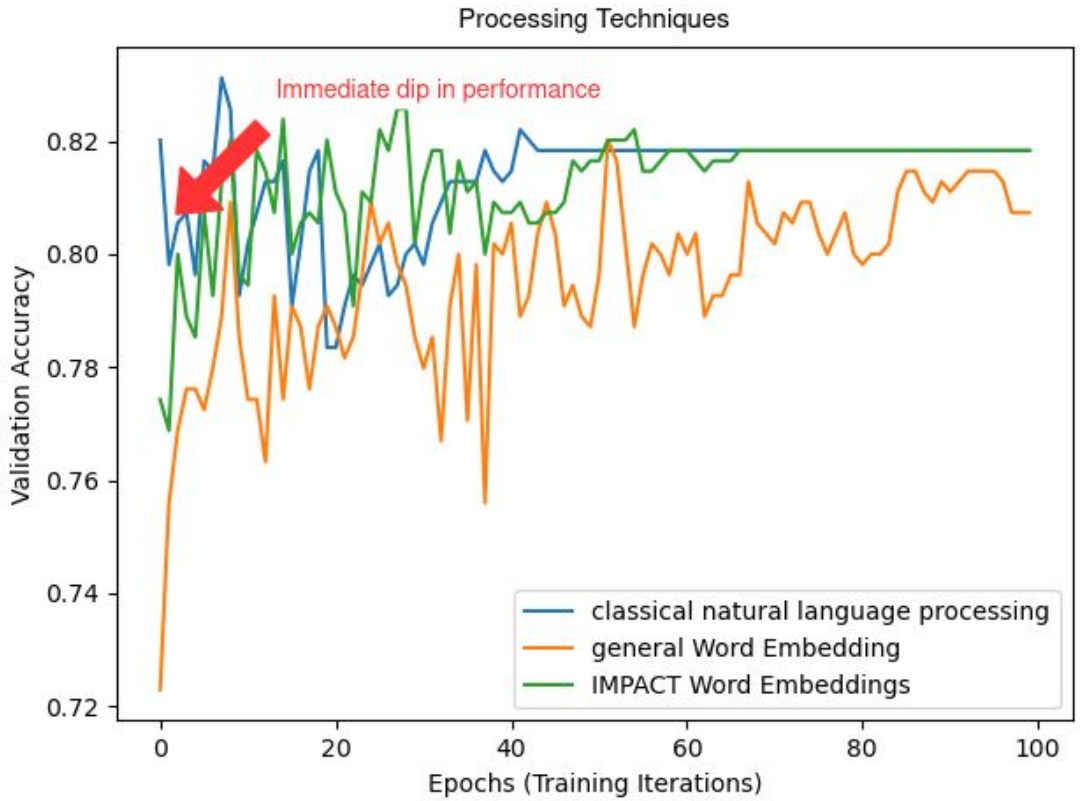
Plot Analysis



Plot Analysis



Plot Analysis



Characteristics of the plots provide insights on our original task.

Original Task

Classify articles into one of three topics:

Fire: 954 datapoints

Sea Level: 954 datapoints

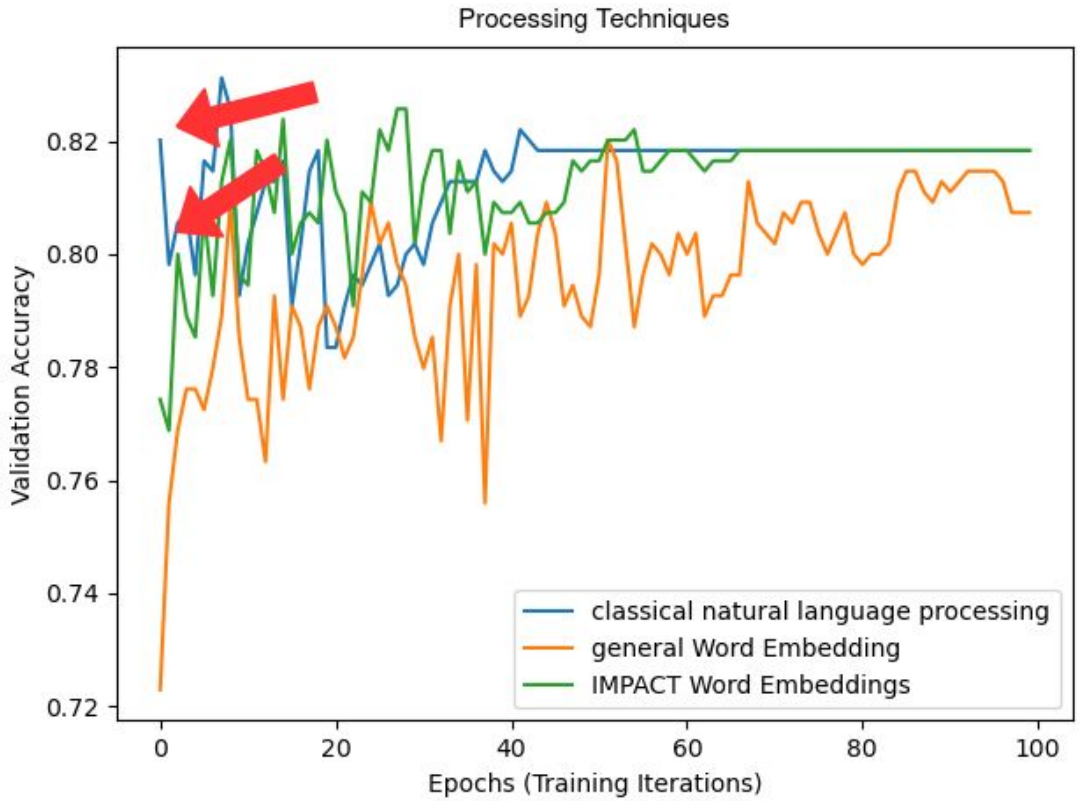
Other: 816 datapoints

Redefined Task

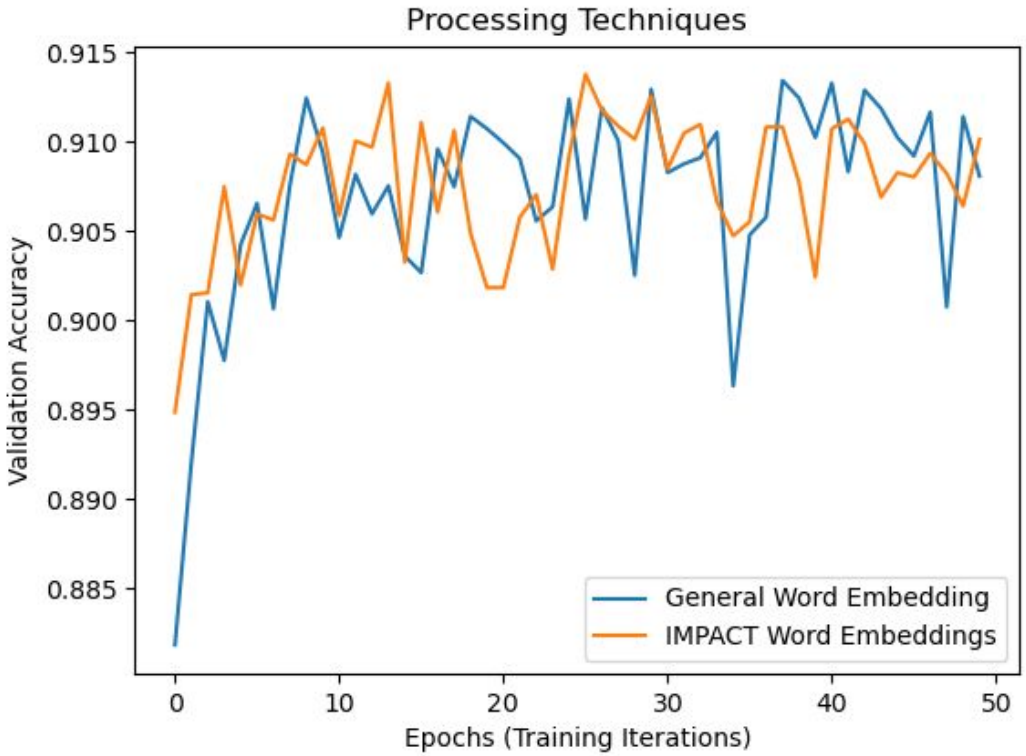
Classify articles on all combinations of four topics:

Count	Fire	Flood	Sea-Level	Snow/Ice
3775	0	0	0	0
3833	0	0	0	1
3807	0	0	1	0
60	0	0	1	1
3768	0	1	0	0
1858	0	1	0	1
3102	0	1	1	0
25	0	1	1	1
3516	1	0	0	0
201	1	0	0	1
62	1	0	1	0
1	1	0	1	1
1352	1	1	0	0
47	1	1	0	1
36	1	1	1	0
2	1	1	1	1

Plot Analysis



Plot Analysis

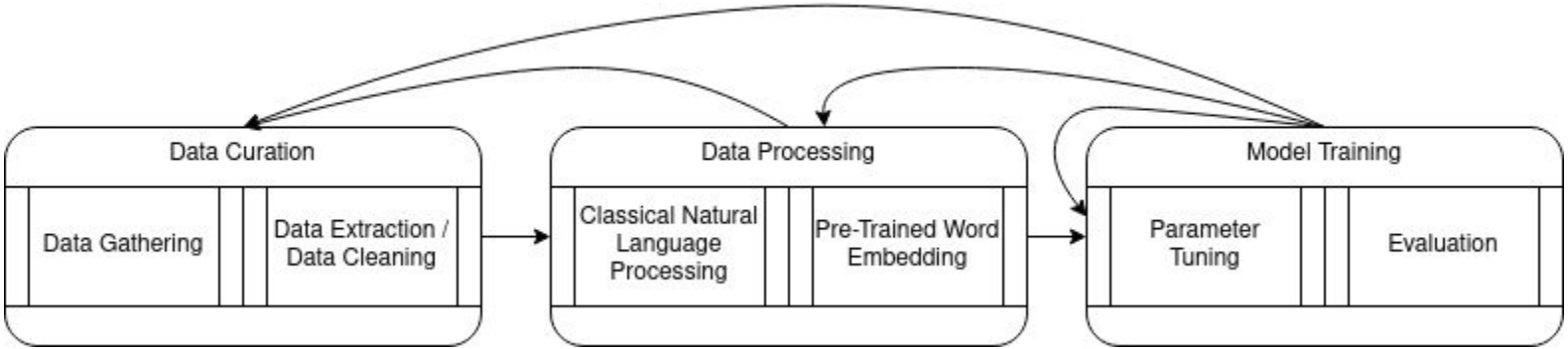


Future Work

- Utilizing full text of papers
- Utilizing references and corresponding text
- Clustering

Takeaways

- The greater machine learning pipeline yields helpful products beyond the algorithm classification results.
- There are a wealth of benefits a machine learning project can gain from modularity.



Backup Slides

Software Used

Word Embeddings

- IMPACT word embeddings
- GloVe (Global Vectors) - https://nlp.stanford.edu/projects/glove/
  - Paper: Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation
  - Data: https://nlp.stanford.edu/data/glove.6B.zip

Python Modules

- NLTK (Natural Language Toolkit) - https://www.nltk.org/
- Scikitlearn - https://scikit-learn.org/stable/
- Pandas - https://pandas.pydata.org/
- NumPy - https://numpy.org/
- PyTorch - https://pytorch.org/
- Matplotlib - https://matplotlib.org/