AON

Analysis & Segmentation of Natural Catastrophe Events

James Lunt

Part 1:

Task 1: Exploratory Data Analysis

Overview: An initial inspection to understand the structure and content of the dataset.

From reviewing the dataset 'Nat Cat Events.csv', 91,479 rows are present with 8 fields, namely:

url	url_mobile	title	seendate	socialimage	domain	language	sourcecountry

The following preliminary analysis tasks are performed:

- Dataset structure summary
- Null count & percentage per column
- Uniqueness percentage per column
- Range & distribution of publication dates
- Title length distribution (characters & words)
- Source country distribution (Top 15)
- URL length distribution
- Domain distribution (Top 30)

Some key findings:

- Title is missing 0.1% values (This will be important for Task 2).
- Title has 71.23% unique values (This will also be important for Task 2).
- Publication dates range from 00:00 January 1st, 2024 – 00:15 January 1st , 2025.
- There is an even distribution of publications throughout the year
- The hour of day with typically the most publications is midnight.
- The average number of words in a title is 12.01 (This will be important for tokenisation in Task 2 and Part 2).

Please see 1_Exploratory_Data_Analysis.ipynb for code and output of this analysis. See plots in Appendix section.

Task 2: Data Cleaning

Overview: Clean the data of articles where the title is not relevant to the following criteria:

- They must contain a location
- They must represent a natural catastrophe event that has occurred

Data Cleaning is broken into four steps:

1. Remove duplicates, null values & whitespaces
2. Find articles containing a location in the title using Named Entity Recognition (NER)
3. Use a Zero-Shot Classifier to capture titles implying a natural catastrophe event
4. Build a Supervised Model to refine less confident Zero-Shot classifications

Details of each step:

Step 1: Remove Duplicates, Null Values & Whitespace

- Use Pandas function to remove duplicate titles, drop any rows with empty titles and strip leading and trailing whitespaces

Step 2: Use NER to find articles containing a location in the title.

- Uses SpaCy (a fast-industrial strength NLP library).
- Selects a SpaCy NER pipeline depending on GPU availability. The pipeline can be set in the configuration file.
  - CPU pipeline en_core_web_sm
  - GPU pipeline en_core_web_trf
- Pipeline structure:

Word Embeddings	POS Tagger	Dependency Parser	NER

- The main difference between the CPU & GPU pipeline is the Word Embeddings. en_core_web_trf is a transformer-based embedding while en_core_web_trf uses tok2vec.
- Locations are found using OntoNotes5 dataset entities:
  - GPE: geopolitical entity like a country, city or state
  - LOC: Non-GPE location like a mountain or body of water
- Dataset saved at this point to titles_containing_locations.csv

Step 3: Use Zero-Shot-Classifier to capture titles implying a natural catastrophe event has occurred

- To label the titles as a natural catastrophe event that has occurred use a Natural Inference Model to decide if a piece of text implies the label:
  - 'natural catastrophe event has occurred'
- A Zero-classification model is downloaded from Hugging Face.
  - In the current execution, facebook/bart-large-mnli is used for optimal accuracy.
  - This model has 400M+ parameters and is not recommended for CPU users, instead use a smaller model such as typeform/distilbert-base-uncased-mnli.
  - The model can be set in the configuration file.
- The model produces prediction propensities that the title implies the label.
- Dataset saved at this point to tiles_zero_shot.csv

Step 4: Build a Supervised Model to refine less confident Zero-Shot classifications.

- Take the rows that the model is extremely confident. Using these titles and propensities, a training set is created where the target variable is:
    - 1 where propensity > 0.99
    - 0 where propensity is <0.01
- The titles are vectorised using TF-IDF
- A logistic regression model is tuned, calibrated and evaluated on the vectorised titles with the following specifications:
    - 80-20 train-test split
    - Hyperparameter tuning of the TF-IDF vectorisation and the logistic regression regularisation parameters.
        - 5-Fold cross-validation during tuning with the training set
    - Calibrate the model with an Isotonic Calibrator.
    - Infer on the test set to produce:
        - ROC plot
        - Classification report
- Given the supervised model performs well on the test set, use this model to infer on the less confident predictions. I.e. predictions a propensity:
    - < 0.99 and > 0.01
- Take the average of the Zero-Shot classification propensity and the supervised model propensity.
- Use a decision threshold on this average propensity to include/exclude titles from the final pre-processed dataset. (In the current execution, decision threshold is 0.85).
- In effect, the supervised model acts as a second opinion to the Zero-Shot-Classification model for less confident predictions.

Final pre-processed dataset saved to: [preprocessed_df.csv]

In the current execution, this data cleaning reduces the dataset size as follows:

Initial Size	Step 1	Step 2	Step 3/4
91,479	65,158	40,989	27,533

Please see [2_Data_Cleaning.ipynb] for code and output of this analysis. See plots in Appendix section.

Part 2:

Overview: Using the pre-processed dataset from Part1, Task 2, categorise each title into 1 of 5 categories.

- Earthquake
- Floods
- Volcano
- Tornado
- Wildfire

For this segmentation. Again, a Zero-Shot-Classification model is used with the same details as in Step 3 in the previous section. Only this time:

- The model is *multi-label* where a probability is given for each category above.
- Each title is then categorised into its label with the highest probability.
- For example:
  Title: 'DSWD DROMIC Report on the Tornado Incident in Brgy . Rizal , Anao , Tarlac , 30 December 2023 , 6PM - Philippines',
  Tornado: 0.984
  Volcano: 0.004
  Floods: 0.003
  Wildfire: 0.003
  Earthquake: 0.003
  Category = Tornado

- Post-analysis is then performed on the categorised date set. Here are some of the key findings:
  - 'Floods' has the most titles
  - Floods and Volcano have the greatest number of uncertain predictions while the other categories have quite confident predictions
  - 'Storm', 'Weather', 'hit', 'New', 'Florida', 'County' are words that appear most frequently among the less confident predictions (Zero-shot probability <0.5 for all categories).

Final segmented dataset saved to segmented_results.csv.

Please see 3_Data_Segmentation.ipynb for code and output of this analysis. See plots in Appendix section.

Appendix

Repository:

james-lunt/Nat_Cat_Events: Analysis & Segmentation of Natural Catastrophe Events

Current Execution spec details:

The Jupyter Notebook code execution uses the following:

- Hardware specs:
  - CUDA device: Quadro P2000 with Max-Q Design
- Software specs:
  - PyTorch version: 2.7.1+cu118
  - A list of requirements to be installed with pip.

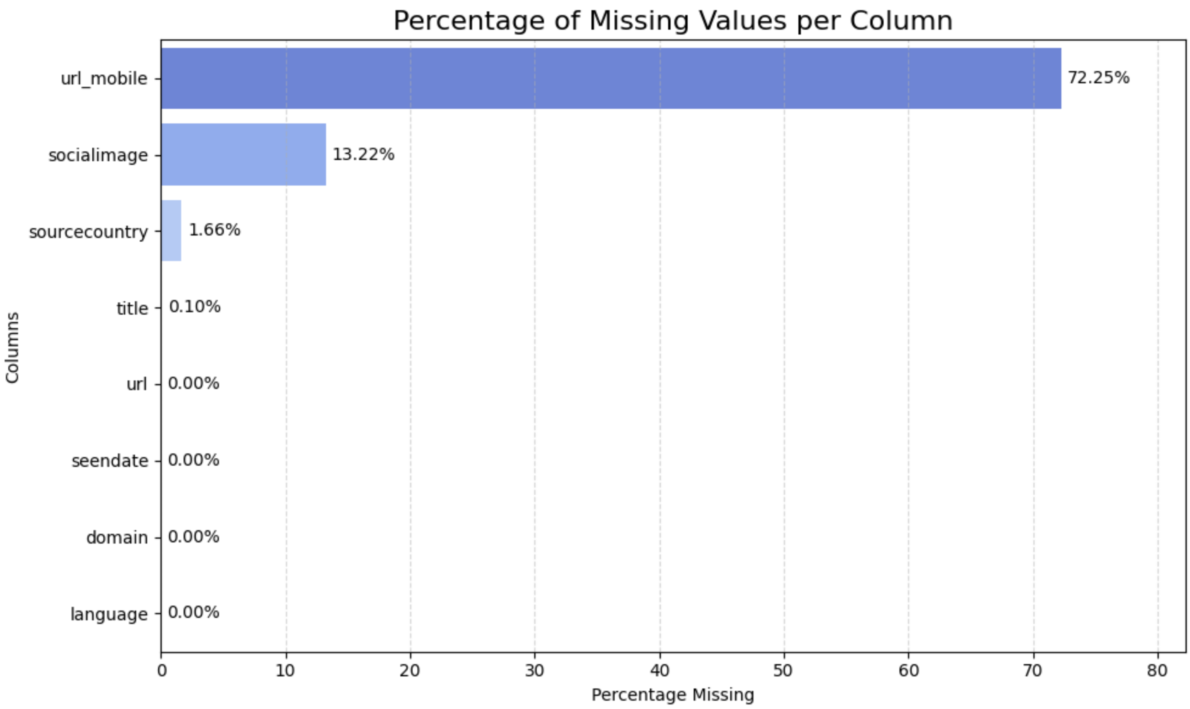
The code can be executed with CPU or GPU availability. See Usage section in ReadMe file for more details.

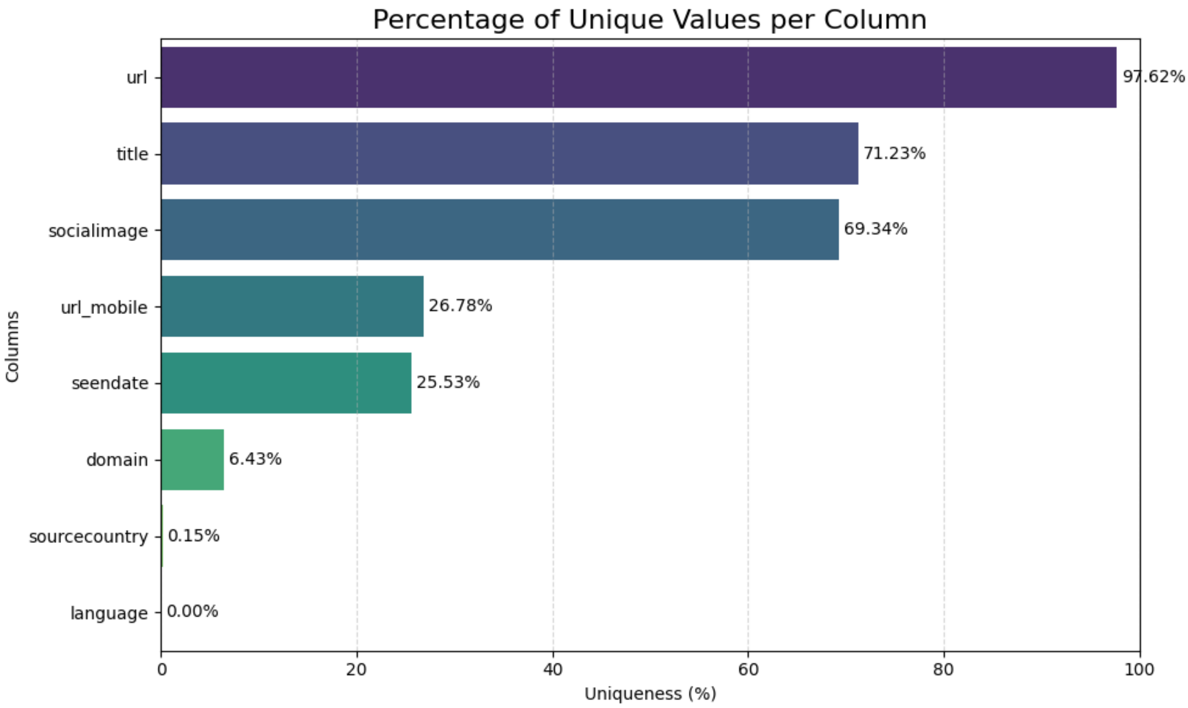
Plots & Figures

Data Analysis

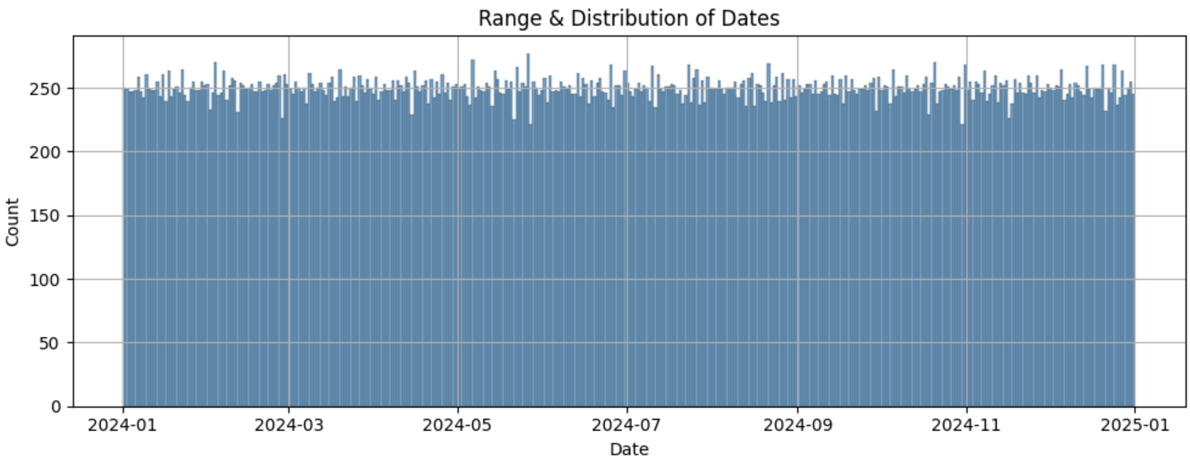
Dataset structure:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 91479 entries, 0 to 91478
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   url            91479 non-null  object
 1   url_mobile     25383 non-null  object
 2   title          91384 non-null  object
 3   seendate       91479 non-null  object
 4   socialimage    79390 non-null  object
 5   domain         91479 non-null  object
 6   language       91479 non-null  object
 7   sourcecountry  89958 non-null  object
dtypes: object(8)
memory usage: 5.6+ MB
```

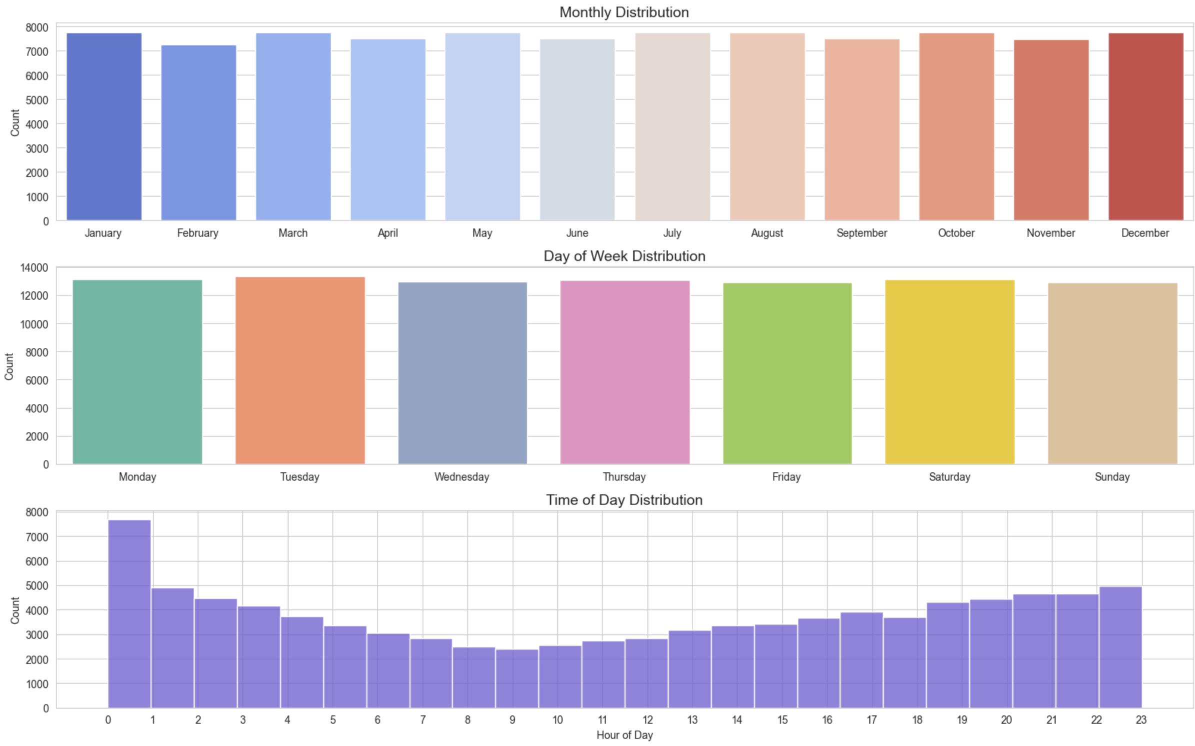


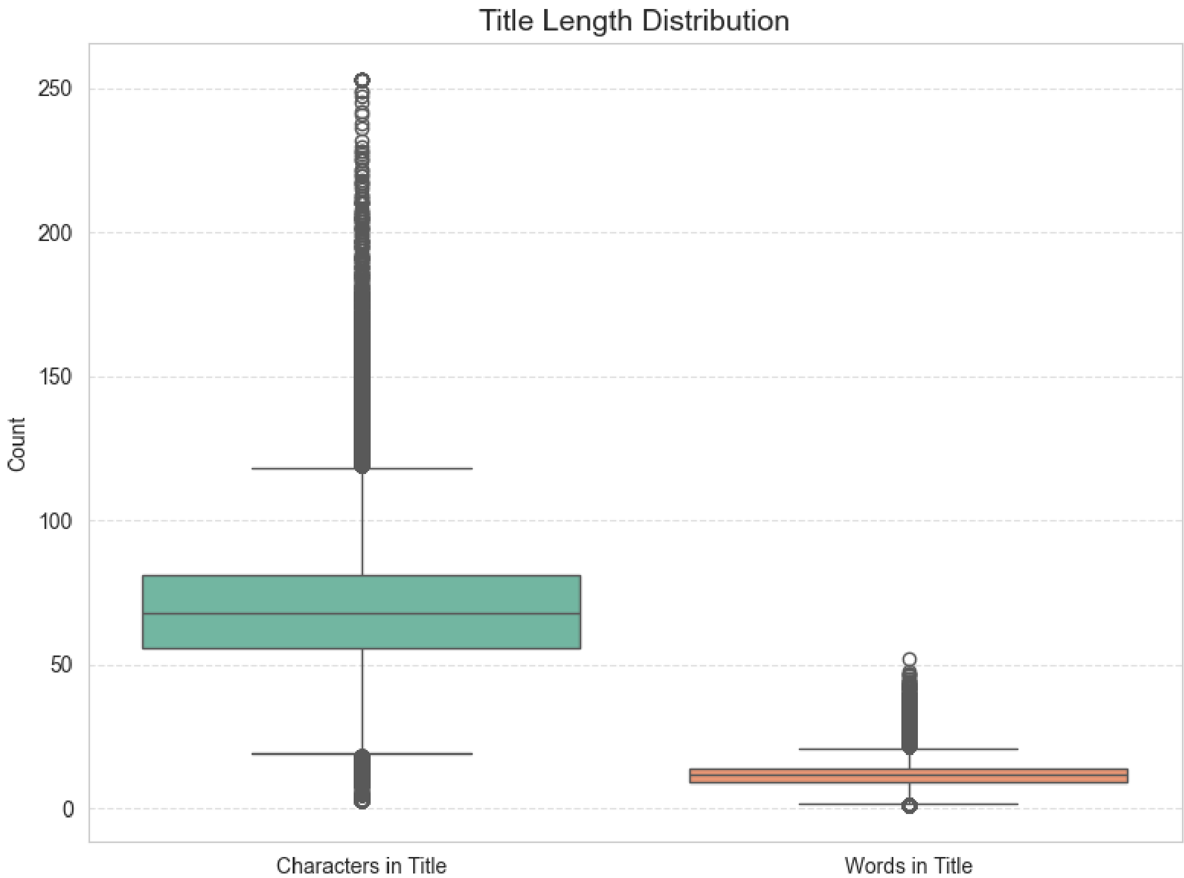


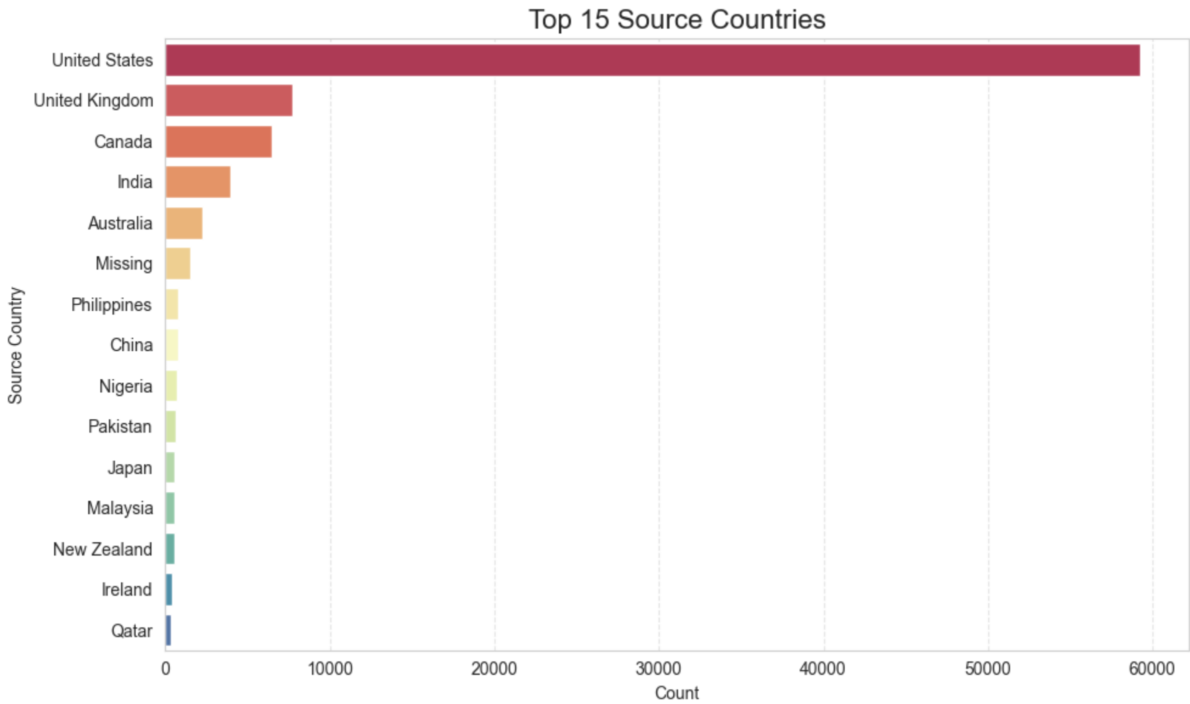
Percentage of Unique Values per Column
url
title
socialimage
url_mobile
seendate
domain
sourcecountry
language
97.62%
71.23%
69.34%
26.78%
25.53%
6.43%
0.15%
0.00%
Columns
0
20
40
60
80
100
Uniqueness (%)

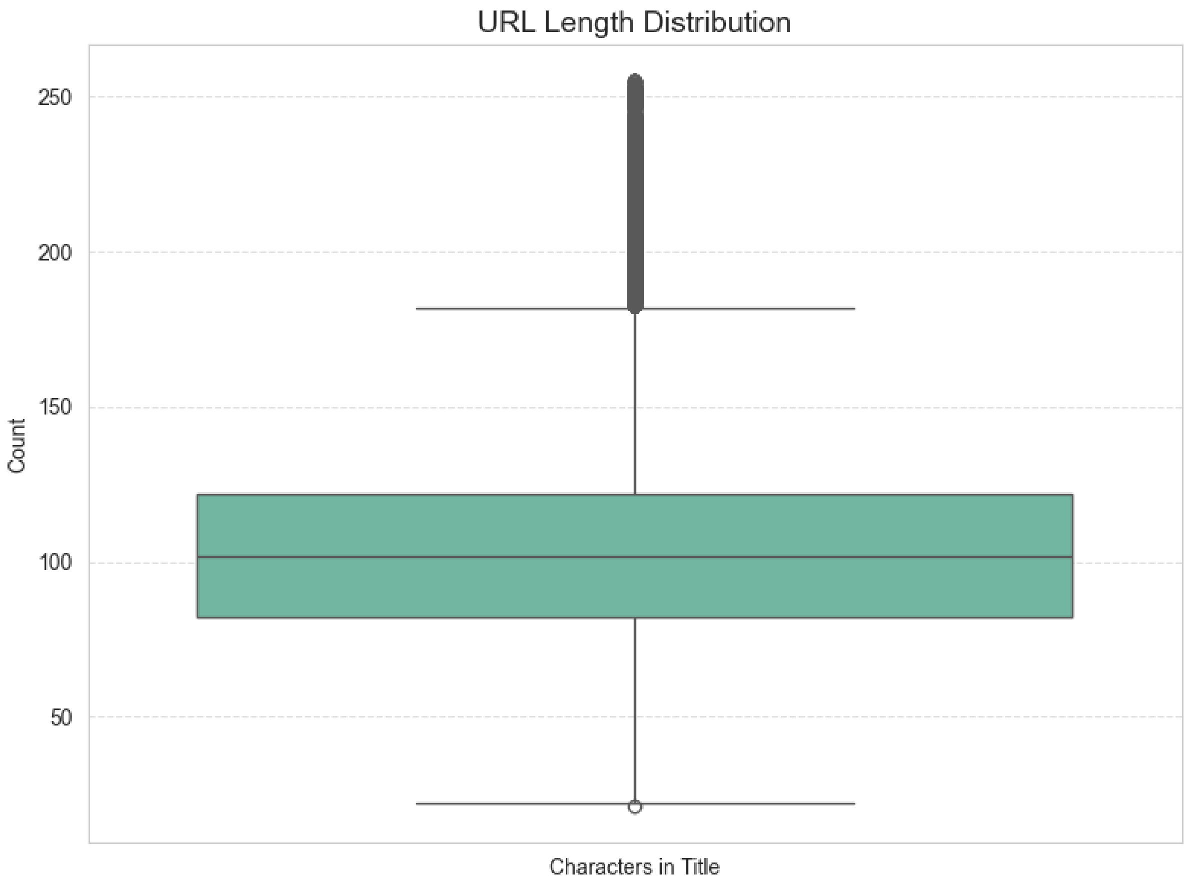


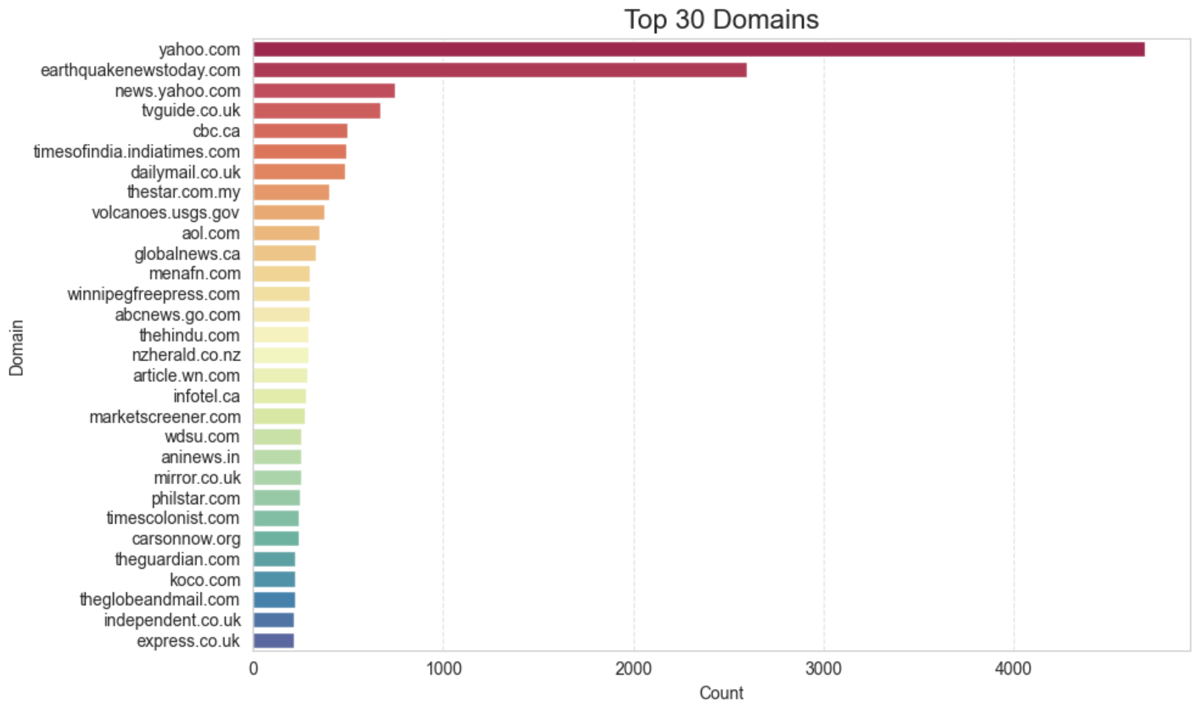
Range & Distribution of Dates
Count
0
50
100
150
200
250
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
Date


Monthly Distribution
Count
January
February
March
April
May
June
July
August
September
October
November
December
Day of Week Distribution
Count
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Time of Day Distribution
Count
Hour of Day


Title Length Distribution
Count
Characters in Title
Words in Title

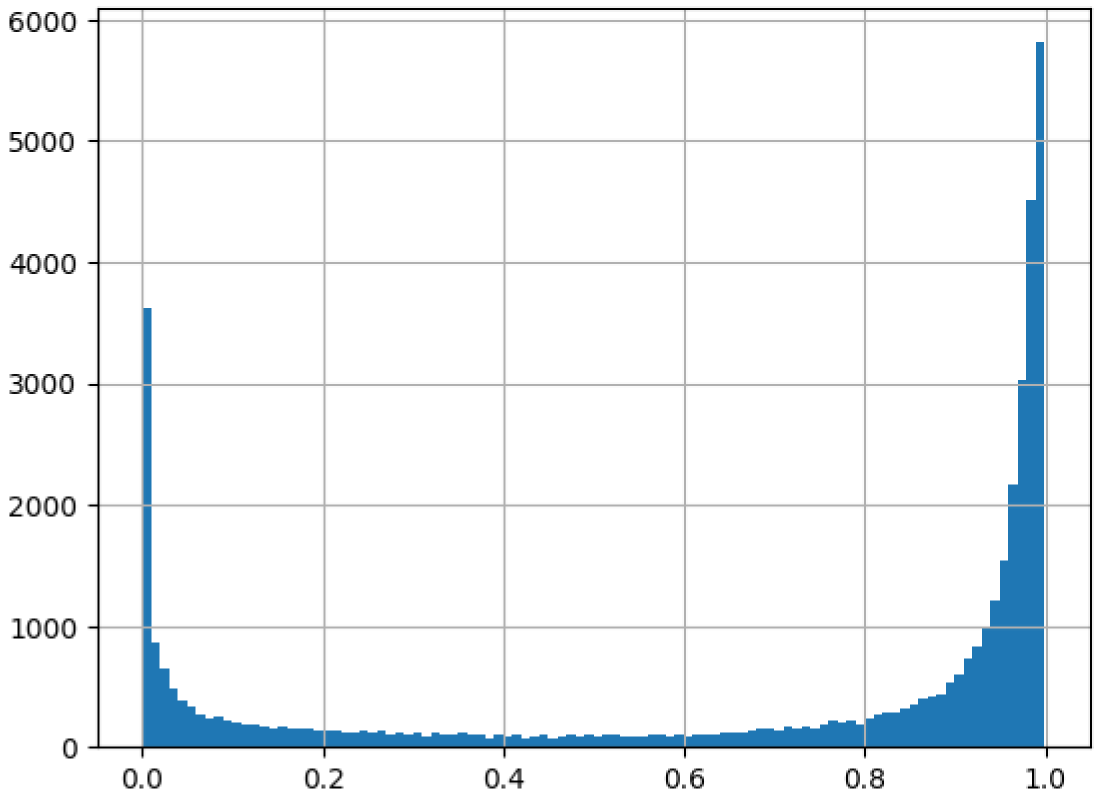

Top 15 Source Countries
United States
United Kingdom
Canada
India
Australia
Missing
Philippines
China
Nigeria
Pakistan
Japan
Malaysia
New Zealand
Ireland
Qatar
Source Country
0
10000
20000
30000
40000
50000
60000
Count


URL Length Distribution
250
200
150
100
50
Count
Characters in Title



Data Cleaning:

Distribution of Zero-Shot-Scores

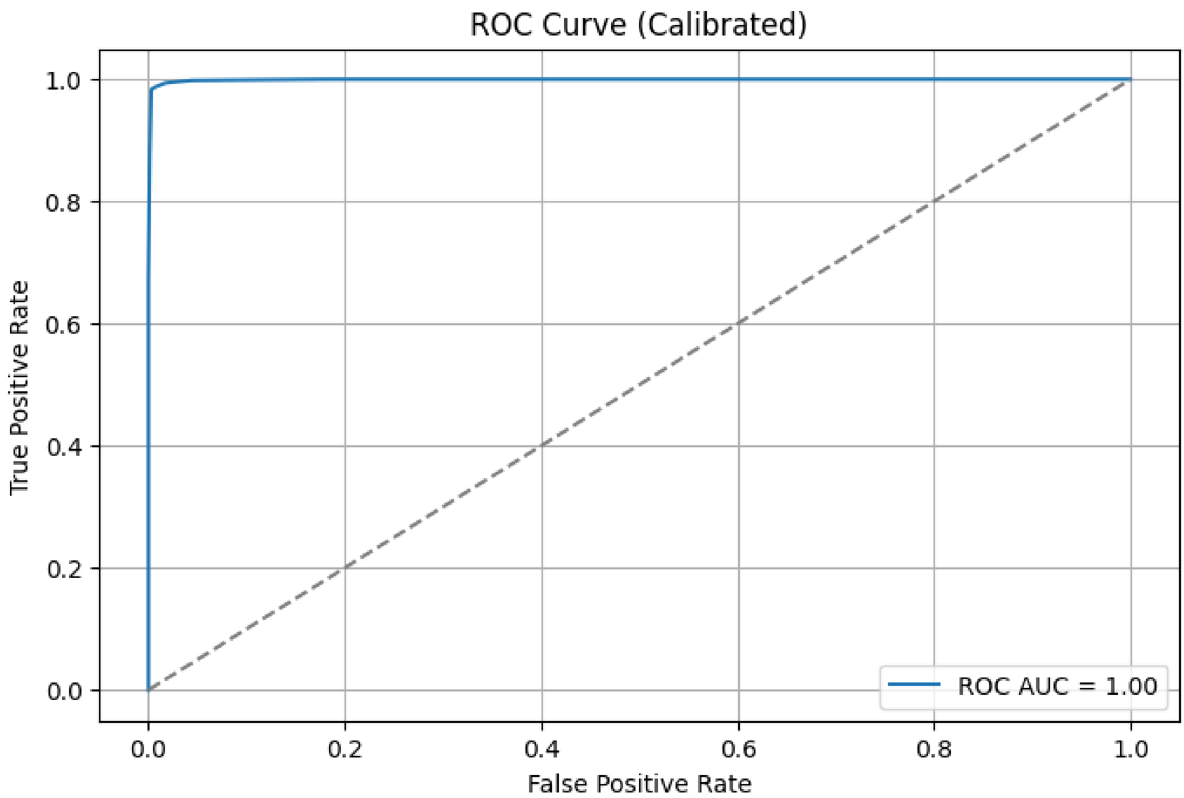


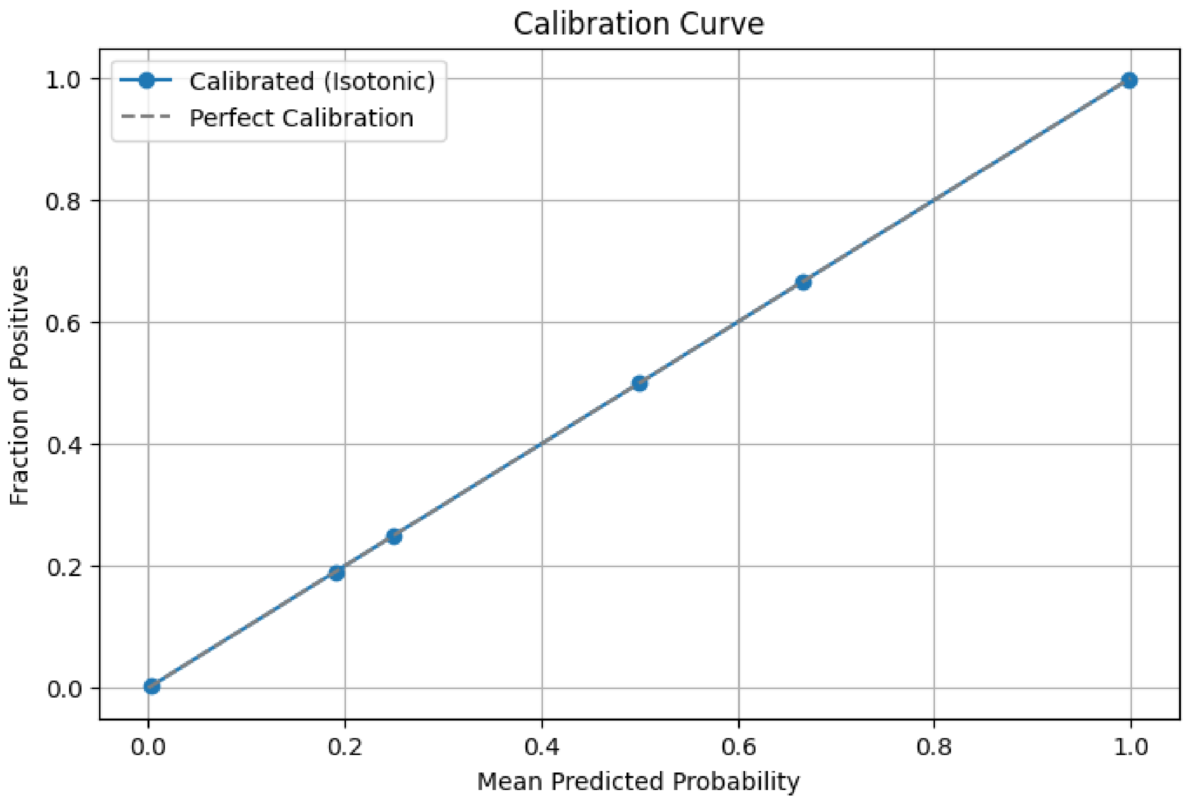
Supervised model best parameters:

Best Parameters: {'clf__C': 10, 'clf__penalty': 'l2', 'tfidf__min_df': 1, 'tfidf__ngram_range': (1, 1)}

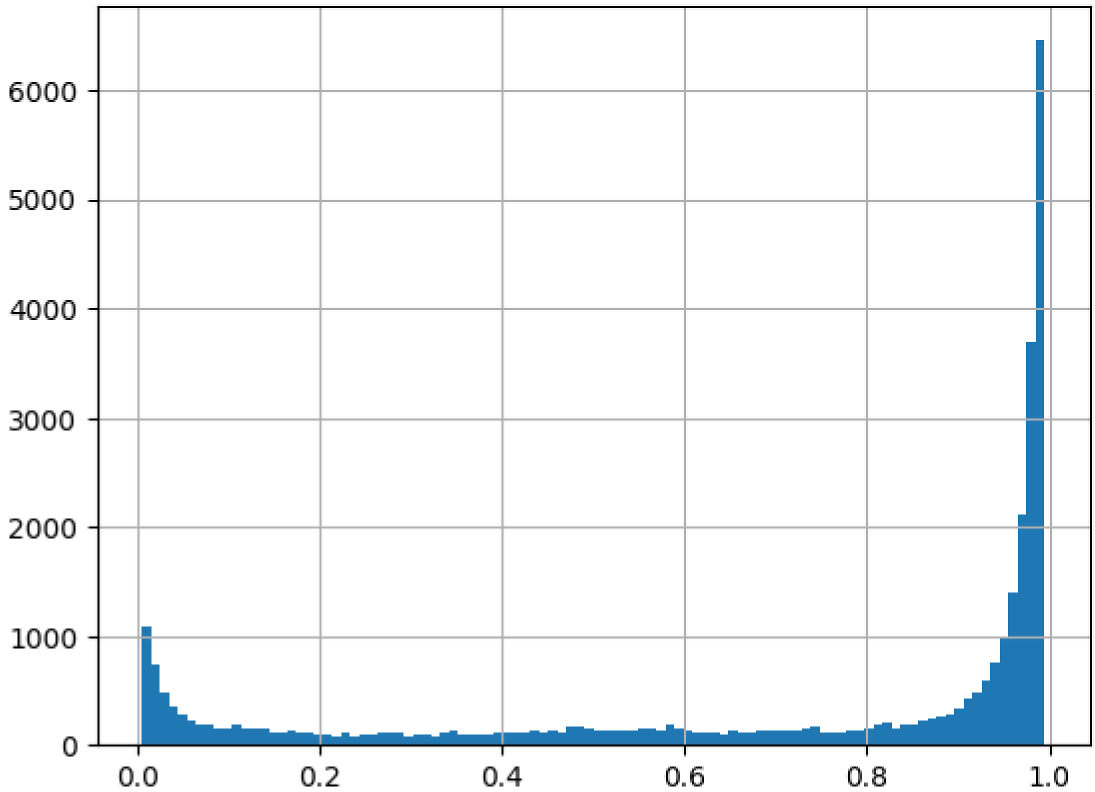
Supervised model classification report:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	722
1	1.00	0.99	0.99	1109
accuracy			0.99	1831
macro avg	0.99	0.99	0.99	1831
weighted avg	0.99	0.99	0.99	1831

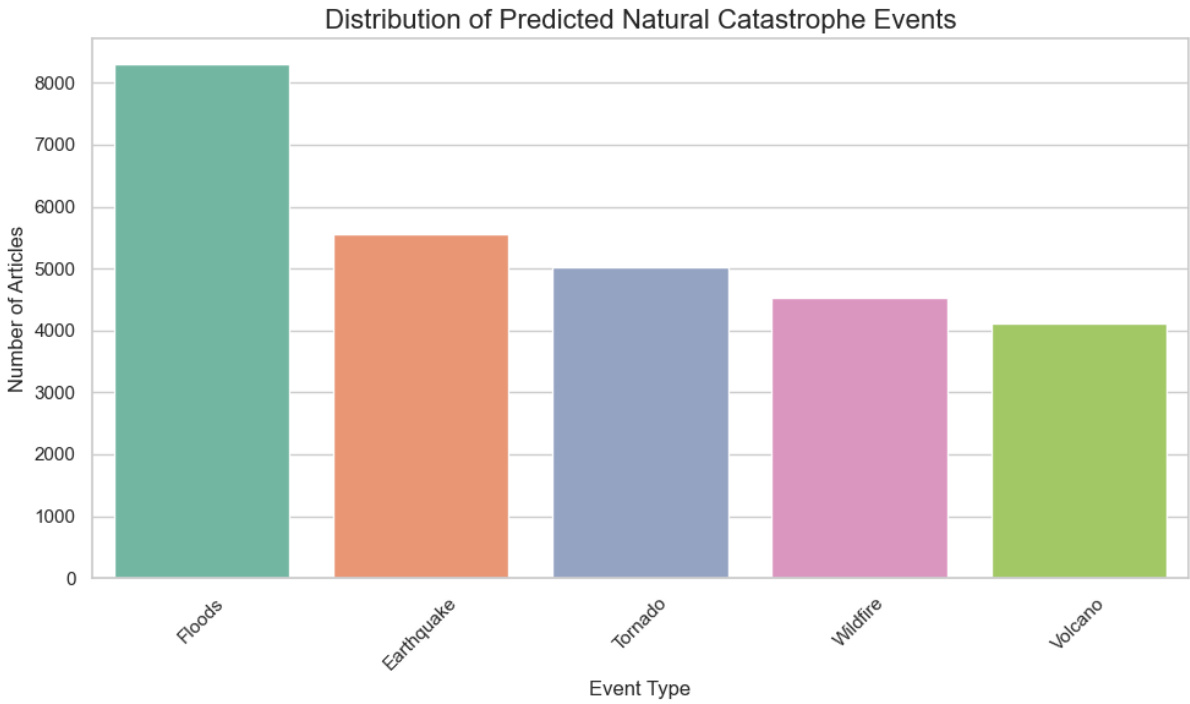


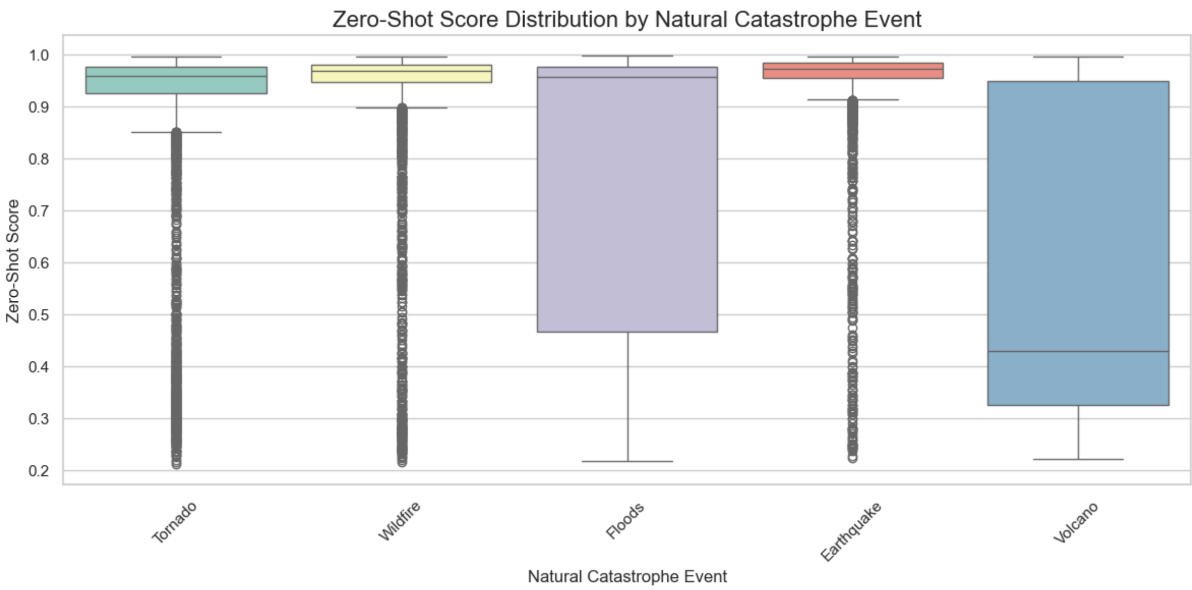


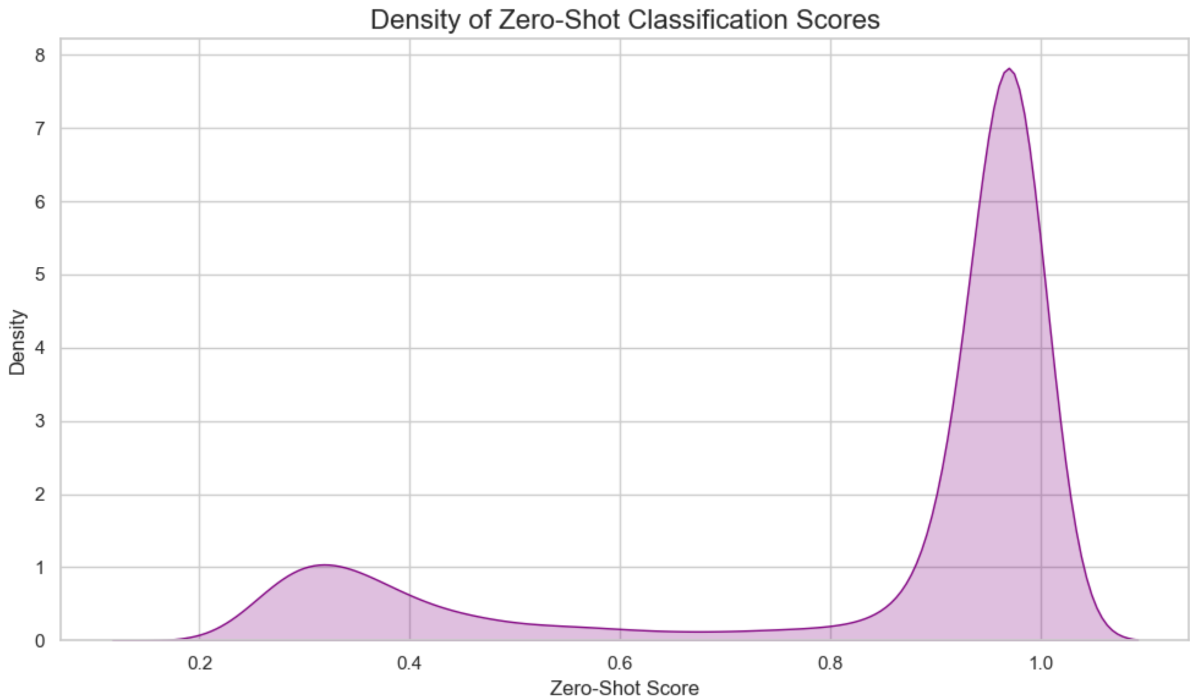
Distribution of zero-shot and supervised model propensities:



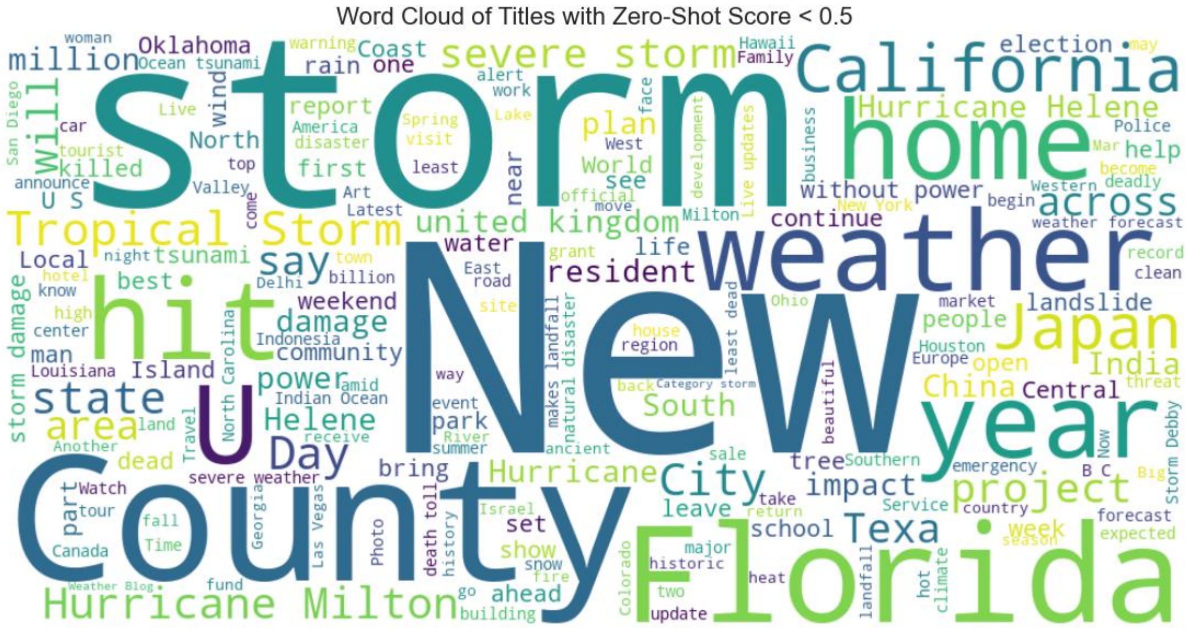
Data Segmentation


Distribution of Predicted Natural Catastrophe Events
Number of Articles
8000
7000
6000
5000
4000
3000
2000
1000
0
Floods
Earthquake
Tornado
Wildfire
Volcano
Event Type


Zero-Shot Score Distribution by Natural Catastrophe Event
Zero-Shot Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Tornado
Wildfire
Floods
Earthquake
Volcano
Natural Catastrophe Event



Density of Zero-Shot Classification Scores
Density
Zero-Shot Score
0
1
2
3
4
5
6
7
8
0.2
0.4
0.6
0.8
1.0



Word Cloud of Titles with Zero-Shot Score < 0.5
Storm
New
County
Florida
weather
home
California
year
hit
Japan
Tropical Storm
Hurricane Milton
Hurricane Helene
severe storm
united kingdom
state
area
Day
Helene
Texa
project
City
impact
resident
across
India
China
Oklahoma
million
Will
killed
North
report
first
plan
World
see
without power
continue
landslide
people
Central
South
open
Hurricane
school
leave
bring
power
community
damage
weekend
water
life
say
tsunami
best
Local
Island
man
dead
storm damage