**GLOBAL ISSUES News Classification Dataset**

**Dataset name: global-issues.csv**

**URL: https://www.kaggle.com/abdulraheemaleem/globalissues**

**Challenges**

The dataset is highly correlated and imbalance in nature. All the classes are equally important so without removing correlated features make a stable classification model is a challenge. Because the dataset is increasing auto on weekly basis using dynamic scrapper

GLOBAL ISSUES is a collection of more than 10k news articles. News articles have been gathered from benchmark news publishers around the world. The dataset is provided by the academic community for research purposes in data mining (clustering, classification, etc), information retrieval (ranking, search, etc), xml, data compression, data streaming, and any other non-commercial activity.

The GLOBAL ISSUES news topic classification dataset is constructed by Abdul Aleem (abdul.raheem.aleem.gcu@gmail.com.edu) and Sohail Asghar. It is constructed to use as a text classification benchmark in the following paper: Abdul, Aleem, Sohail Asghar. GLOBAL ISSUES: A first public new benchmark dataset from benchmark news publishers on GLOBAL ISSUES around the world (submitted in journal) 2021  
This dataset is generating real-time so this type of dataset is not available on popular benchmark websites like Kaggle, UCI, and other free dataset repositories that help out researchers to find the latest trends, impacts on society regarding GLOBAL ISSUES and did classification, clustering, and other text mining task.

**DESCRIPTION**

The GLOBAL ISSUES news classification dataset is constructed by choosing 13 largest classes from the original corpus. The total number of samples is 16000 and will increase weekly basis.

**Content**

This dataset has 13 classes, The classes of this data set are flowing:  
1 global energy politics,  
2 international security  
3 nuclear security  
4 human rights  
5 Palestine conflicts  
6 middle east crisis  
7 climate change  
8 Kashmir Issue  
9 COVID 19  
10 global economy  
11 world population trends  
12 nuclear politics in South Asia  
13 international trade

## Train Data Set

Dataset:News\_Category\_Dataset\_v2.json

url: https://www.kaggle.com/rmisra/news-category-dataset

**Context**

This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from [HuffPost](https://www.huffingtonpost.com/). The model trained on this dataset could be used to identify tags for untracked news articles or to identify the type of language used in different news articles.

**Content**

Each news headline has a corresponding category. Categories and corresponding article counts are as follows:

* POLITICS: 32739
* WELLNESS: 17827
* ENTERTAINMENT: 16058
* TRAVEL: 9887
* STYLE & BEAUTY: 9649
* PARENTING: 8677
* HEALTHY LIVING: 6694
* QUEER VOICES: 6314
* FOOD & DRINK: 6226
* BUSINESS: 5937
* COMEDY: 5175
* SPORTS: 4884
* BLACK VOICES: 4528
* HOME & LIVING: 4195
* PARENTS: 3955
* THE WORLDPOST: 3664
* WEDDINGS: 3651
* WOMEN: 3490
* IMPACT: 3459
* DIVORCE: 3426
* CRIME: 3405
* MEDIA: 2815
* WEIRD NEWS: 2670
* GREEN: 2622
* WORLDPOST: 2579
* RELIGION: 2556
* STYLE: 2254
* SCIENCE: 2178
* WORLD NEWS: 2177
* TASTE: 2096
* TECH: 2082
* MONEY: 1707
* ARTS: 1509
* FIFTY: 1401
* GOOD NEWS: 1398
* ARTS & CULTURE: 1339
* ENVIRONMENT: 1323
* COLLEGE: 1144
* LATINO VOICES: 1129
* CULTURE & ARTS: 1030
* EDUCATION: 1004

**Acknowledgements**

This dataset was collected from [HuffPost](https://www.huffingtonpost.com/).

**Inspiration**

* Can you categorize news articles based on their headlines and short descriptions?
* Do news articles from different categories have different writing styles?
* A classifier trained on this dataset could be used on a free text to identify the type of language being used.

**Want to contribute your own datasets?**

If you are interested in learning how to collect high-quality datasets for various ML tasks and the overall importance of data in the ML ecosystem, consider reading my book [Sculpting Data for ML](http://amzn.com/B08RN47C5T).

**Citation**

If you're using this dataset for research purposes, please use the following BibTex for citations:

@dataset{dataset,

author = {Misra, Rishabh},

year = {2018},

month = {06},

pages = {},

title = {News Category Dataset},

doi = {10.13140/RG.2.2.20331.18729}

}

@book{book,

author = {Misra, Rishabh and Grover, Jigyasa},

year = {2021},

month = {01},

pages = {},

title = {Sculpting Data for ML: The first act of Machine Learning},

isbn = {978-0-578-83125-1}

}

Please link to [rishabhmisra.github.io/publications](https://rishabhmisra.github.io/publications/) in your report.

Thanks!

**Other datasets**

Please also checkout the following datasets collected by me:

* [News Headlines Dataset For Sarcasm Detection](https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection)
* [Clothing Fit Dataset for Size Recommendation](https://www.kaggle.com/rmisra/clothing-fit-dataset-for-size-recommendation)
* [IMDB Spoiler Dataset](https://www.kaggle.com/rmisra/imdb-spoiler-dataset)